Gait Recognition using Pose Estimation and Signal Processing

Vítor Cézar de Lima and William Robson Schwartz*

Smart Sense Laboratory, Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{vitorcezar,william}@dcc.ufmg.br

Abstract. Gait is a biometry that differentiates individuals by the way they walk. Research on this topic has gained evidence since it is unobtrusive and can be collected at distance, which is desirable in surveillance scenarios. Most of the previous works have focused on human silhouette as representation. However, they suffer from many factors such as movement on scene, clothing and carrying conditions. To avoid such problems, this work employs pose estimation to retrieve the coordinates of body parts, which are transformed into signals and movement histograms to be used as feature descriptors. While the former descriptors are used with the Subsequence Dynamic Time Warping that compares signals from probe and gallery, the Euclidean distance is used on the latter to find the person on gallery that is closest to probe. Finally, the outputs of both are fused. This work was evaluated on all views of CASIA Dataset A and compared to existing ones, demonstrating its efficacy.

Keywords: Gait Recognition · Biometry · Computer Vision.

1 Introduction

Gait is a biometry employed to identify individuals by the way they walk [7]. It has gained evidence in recent years due to its advantages comparing with other biometrics (e.g., iris, face and fingerprint), since it can be collected at distance and it does not need the subject cooperation.

Works on gait recognition can be divided in two main approaches: model-based and model-free. While the former models characteristics of human motion pattern, the latter employs other features to represent gait, such as silhouette [22], symmetry of human motion [5] and Fourrier descriptors [14].

An example of a model-based approach is the work of Wang et al. [20], that uses as dynamic feature a tracking operation that calculates joint-angle trajectories of the main lower limbs. Another work is Wagg et al. [19], where anatomical data is used to generate shape models consistent with regular human body proportions and create prototype gait motion models adapted to fit each subject. The problem is that model-based methods present a high computational cost since they need to model human motion patterns.

* Corresponding author

The methods of model-free approach are simple and present lower computational requirements. They usually use the silhouette to represent a subject, such as the gait features [10, 12, 16], where the gait energy image [13] is the most influential, in which silhouettes from a gait cycle are normalized and aligned and then, have their mean values calculated and used as template. However, this approach is severely effected by clothing, carried objects and the camera view of the silhouette.

There are works that address the problems related to camera view, clothing and carrying conditions. Liu and Zheng [1] developed a cross-view approach that does not require prior information regarding the probe angle. It uses Gaussian process for view angle estimation and correlation strength from canonical correlation analysis to perform gait recognition. The algorithm was tested on CASIA Dataset B [23] and performed better than other existing works. However, the results still need to be improved, specially when other conditions are tested, such as carrying and clothing. Isaac et al. [6] proposed a genetic template segmentation to select silhouette parts to be used for classification. For each angle, the genetic algorithm automates the boundary positions and selects the parts to use, from which a view-estimator is able to determine the probe angle and selects the suitable view-specific classifier to recognize.

Despite the advances on gait recognition with methods based on silhouettes, there are intrinsic limitations on such representation. Silhouettes are not robust on outdoor scenes, due to movements of people and objects. In addition, as Sarkar et al. [18] pointed out, they change on different surface types and silhouettes from the same person obtained at different days are hard to be recognized. Therefore, due to these limitations, we employ pose estimation to obtain the coordinates of the body parts [2]. This information is processed generating signals and movement histograms, which are used as features, robust to the aforementioned aspects. Two methods are employed to perform the recognition. While the first uses Subsequence Dynamic Time Warping to compare signals from probe and gallery, ranking based on minimum matching distance cost, the second computes the Euclidean distance of the movement histograms to define the person from gallery that is closest to the probe. Finally, a score fusion is used on the result of these two methods to provide the identity of a subject.

Our main contributions are the following. The creation of a gait representation that is robust to clothing and carrying conditions and can be used on all views and the development of two signal processing methods based on the pose estimation that are efficient on gait discrimination among individuals.

We evaluate the proposed approach on all views of the CASIA Dataset A [22]. Even though our proposed approach neither uses machine learning techniques nor deep learning-based approaches, we are able to achieve recognition accuracy above 92.5% on all views. The results are compared to state-of-art works, showing that our approach achieves similar accuracy to the best works on lateral and oblique views and the same accuracy of the best work on the frontal view.

2 Feature Extraction

To extract the body coordinates, we use the pose estimator proposed by Cao et al. [2]. Their method returns $P^i_{b,t}$ for each frame t on gait sequence i and body part indexed by b. $1 \leq t \leq T^i$, where T^i is the total number of frames on sequence i. $P^i_{b,t}$ is a coordinate (x, y) and x and y values are referenced by $P^i_{b,t}.x$ and $P^i_{b,t}.y$, respectivelly. The coordinates have non-negative values, except when the body part is not found due to occlusion, when they have the invalid values $(-1, -1)$.

The coordinates returned by the pose estimator are from 18 body parts: neck, nose and both ears, wrists, elbows, hips, knees, ankles, shoulders and eyes. While ears, eyes and nose are not considered because their positions do not help on the gait recognition, the remaining 13 body parts, indexed by b ranging from 1 to 13, are used. For instance, the index b for the neck is 1 and $P^i_{1,t}$ is the neck coordinate at the t-th frame and in the i-th sequence.

The invalid coordinates generated by the pose estimator due to occlusion may interfere with the results. So, to avoid interference, a tracking of invalid body parts is performed, creating the noise indexing N^i for sequence i that saves the indexes of all body parts whose percentage of invalid coordinates is higher than a defined threshold. It is used on classification to eliminate "noisy" body parts on the Subsequence Dynamic Time Warping and the Euclidean distance calculation (see Sections 3.1 and 3.2). N^i is defined as

$$N^i = \left\{ b : \frac{\#\{P^i_{b,t}.x = -1 \ \ \forall t \in [1, 2, ..., T^i]\}}{T^i} > \gamma \right\}, \tag{1}$$

where γ is a parameter that represents the noise tolerance. Therefore, it should be defined in a way that increases recognition accuracy, and this is done on Section 4.

2.1 Feature based on Body Part Signals

Body part signals are used to get dynamic information on gait sequence, differentiating individuals by the way their body locations vary on time relative to neck position.

Signals to represent gait movement from sequence i are created, represented by S^i. Each b from 2 to 13 (the neck is used on the formula, but signals for it are not created because they would have only zeros) will generate two lines on S^i: one for its x coordinate value and the other for y. S^i has 24 lines and T^i columns and it is obtained using

$$S^i_{2b-3,t} = \begin{cases} -1, & \text{if } P^i_{b,t}.x = -1 \\ \frac{P^i_{b,t}.x - P^i_{1,t}.x}{max_j P^i_{j,t}.y - P^i_{1,t}.y}, & \text{otherwise} \end{cases} \tag{2}$$

$$S^i_{2b-2,t} = \begin{cases} -1, & \text{if } P^i_{b,t}.y = -1 \\ \frac{P^i_{b,t}.y - P^i_{1,t}.y}{max_j P^i_{j,t}.y - P^i_{1,t}.y}, & \text{otherwise} \end{cases} \tag{3}$$

According to the equations, the person position on frame is not considered because the distances are relative to neck position. The denominator of S^i is the vertical distance of the neck to one of the feet (one foot is always on the floor, so maximum y is from it), making the signals invariant to the person distance on the video.

After the signal creation, each line of S^i has its invalid values removed using linear interpolation and median filter is also applied. Figure 1 shows examples of signals from all views, where some periodicity can be observed.

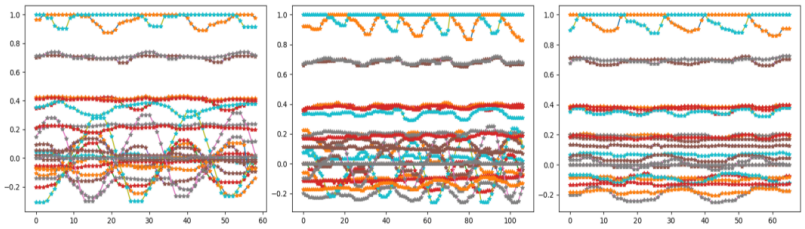

Fig. 1. Examples of signals from lateral, oblique and frontal view, respectively. The lines correspond to x and y distance of body parts to the neck position on coordinates. The distances vary with time, changing the amplitude.

2.2 Feature based on Movement Histograms

The movement histograms capture static gait information, dividing the values of body part signals on intervals to represent a gait sequence.

After the signals creation and processing as described in the previous section, the movement histogram H^i is created for sequence i. S^i values are divided on intervals in $(-1, 1]$, and each occurrence of a value in S^i increments its corresponding position on H^i. Then, each value on H^i is divided by T^i, turning the histogram values invariant to sequence size. This is represented by

$$H^i_{l,j} = \frac{\#\left\{\left\lceil \frac{nVals(S^i_{l,t}+1)}{2} \right\rceil = j \;\; \forall t \in [1, 2, ..., T^i]\right\}}{T^i}, \tag{4}$$

for $1 \leq l \leq 24$ and $1 \leq j \leq nVals$, where $nVals$ is the number of intervals in $(-1, 1]$.

Increasing $nVals$ make easier to differentiate individuals. The problem is that if $nVals$ is extremely high, the division of values of S^i on H^i will be sparse and the recognition will be affected. Therefore, it is necessary to find an optimum value of $nVals$ that can differentiate individuals without decreasing recognition (see experimental results section). Figure 2 shows movement histograms for different values of $nVals$.

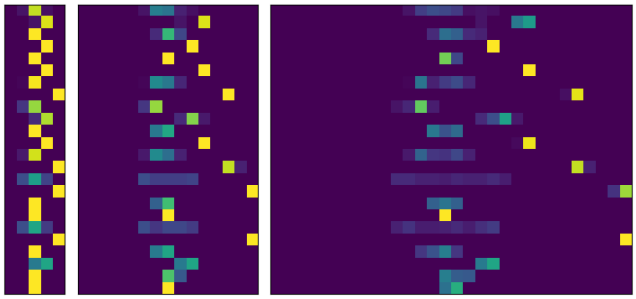

Fig. 2. Histograms with $nVals$ equal to 5, 15 and 30, respectively. Each line is related to one body part and each column corresponds to an interval on histogram. The images show that increasing $nVals$ increases the distribution of signal values on histogram, but causes sparsity on some intervals.

3 Gait Recognition

This section present two methods used for recognition and their fusion. They use the features of body part signals and movement histograms presented on the last section.

In the recognition stage, each person g from gallery is compared to probe p. The goal is to find the person g that minimizes the cost function $D^{p,g}_{sdtw}$, $D^{p,g}_{dist}$ or $D^{p,g}_{fusion}$.

The union operation is applied on the indexes of body parts on N^p and N^g, creating $N^{p,g}$. S^{i*}, S^{p*}, H^{i*} and H^{p*} are created from S^i, S^p, H^i and H^p, removing lines corresponding to body parts indexed on $N^{p,g}$.

The following subsection presents how $D^{p,g}_{sdtw}$ is calculated using Subsequence Dynamic Time Warping. Then, Section 3.2 describes the use of the Euclidean distance with the movement histograms to calculate $D^{p,g}_{dist}$. Finally, the last section discusses the fusion $D^{p,g}_{fusion}$ of the results.

3.1 Subsequence Dynamic Time Warping

Dynamic Time Warping (DTW) [17] is a nonlinear dynamic programming time normalization technique, where a warping function is computed to map the time axis from the probe to gallery samples. Its advantages are that signals which are compared do not need to have the same number of samples and this technique is robust to changes in walking speed [8]. It is used on gait applications for signal normalization [4, 20] and matching [8]. In the latter case, gait sequences that have low minimum matching cost are more likely to be from the same person.

On this work, signal matching is applied, using a variation of DTW called Subsequence Dynamic Time Warping (SDTW) to find S^{g*} from the gallery where a subsequence within S^{p*} is optimally fitted using squared distance. This subsequence is composed of 26 consecutive columns, which is the size of a gait cycle.

The result of SDTW is also normalized by the number of lines of S^{g*}. This operation is important, giving the variation on number of lines of S^{g*} for different persons from gallery, because of N^g information. The cost function of SDTW is defined in the equation below:

$$D^{p,g}_{sdtw} = \frac{SDTW(S^{p*}_{:,1:26}, S^{g*})}{\sqrt{num\_lines(S^{g*})}} \tag{5}$$

3.2 Euclidean Distance

To recognize gait using movement histograms, Euclidean distance is applied comparing probe H^{p*} with H^{g*} from gallery. These features have two dimensions, so they are vectorized for comparison. The Euclidean distance used on movement histograms is defined in the equation below:

$$D^{p,g}_{dist} = \frac{\|vec(H^{p*}) - vec(H^{g*})\|}{\sqrt{num\_lines(H^{g*})}} \tag{6}$$

The distance is used to rank the individuals from gallery based on their distance from probe. The result is normalized according to the the number of lines on H^{g*} for the same reason discussed in the previous subsection.

3.3 Score Fusion

Fusion of features is a common operation used to improve recognition accuracy. In this work it is done using information from SDTW and Euclidean distance. Score fusion is applied, combining results of the two methods as

$$D^{p,g}_{fusion} = \alpha D^{p,g}_{dist} + (1-\alpha) D^{p,g}_{sdtw}, \tag{7}$$

$$\text{with } 0 \leq \alpha \leq 1.$$

In this equation, increasing α favors the results of Euclidean distance and decreasing α favors SDTW. An experiment described in the next section estimates the best value of α in which the fusion leads to better gait recognition.

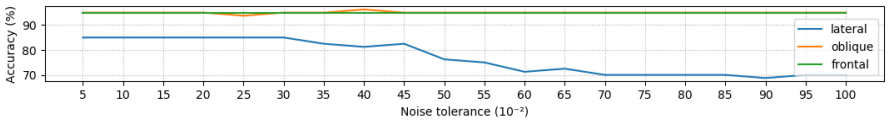

Fig. 3. Accuracy on D_{sdtw} varying the noise tolerance γ. The best results are found when γ is between 0.05 to 0.2.

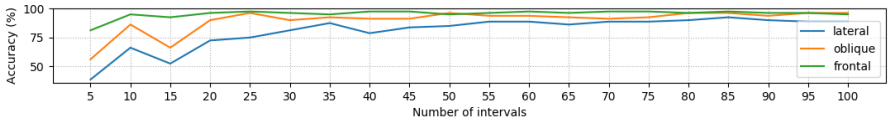

Fig. 4. Accuracy on D_{dist} varying the number of intervals $nVals$. The value 85 gives the best results.

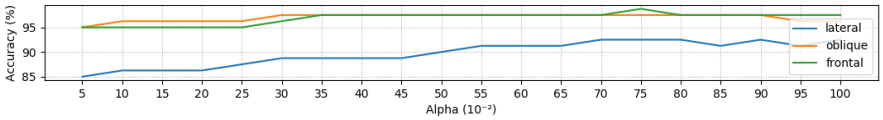

Fig. 5. Accuracy on D_{fusion} varying the score fusion weight α. The value 0.75 gives the best results.

4 Experimental Results

Experiments are conducted on the CASIA Dataset A [22], in which data were captured at 25 FPS and has frame resolution of 352 × 240 pixels. This dataset has sequences from lateral, oblique and frontal views. There are 20 subjects and they have four gait sequences for each view angle. Two of these sequences have the same walking direction and on the other half the walking direction is reversed. For each pair of videos with the same configuration of person, angle and movement direction, one is used as gallery and the other as probe.

The next paragraphs describe and analyze the feature, classification parameters and the results achieved are compared to other methods in the literature.

Noise tolerance. The noise tolerance γ defines which body parts will not be used with SDTW and Euclidean distance because of occlusion. This parameter is tested on the SDTW method and was varied from 0.05 to 1. According to the results showed in Figure 3, occlusion has a great impact on recognition and the best results were found when γ is between 0.05 and 0.2. Giving that lower values remove more body parts on calculations and make the computation simpler as consequence, 0.05 has been chosen to be used in the remaining experiments.

Number of intervals. This experiment evaluates the number of intervals $nVals$ on the movement histogram. It is responsible for the size of intervals, which are used to distinguish gait on sequences. $nVals$ was varied from 5 to 100. According to Figure 4, the best results are achieved by $nVals$ equals to 85, which maximizes accuracy on D_{dist}. This value is used in the remaining experiments.

Score fusion weight. The score weight α is used to fuse results from SDTW and Euclidean distance. It ranges from 0 to 1 and, when α presents higher values, it favors Euclidean distance and when it has lower values, SDTW is favored. In our experiment, we varied α from 0.05 to 1. According to Figure 5, the best

Table 1. Rank 1 recognition on CASIA Dataset A.

Work/view	Lateral	Oblique	Frontal
Zhang et al. [24]	55%	80%	85%
Wang et al. [21]	88.75%	87.5%	90%
Liu et al. [11]	85%	87.5%	95%
Chen et al. [3]	90%	75%	-
Wang et al. [20]	97.5%	-	-
Nizami et al. [15]	**100%**	-	-
Kusakunniran et al. [9]	**100%**	**100%**	**98.75%**
SDTW method	85%	95%	95%
Euclidean method	92.5%	96.25%	97.5%
Fusion of methods	92.5%	97.5%	**98.75%**

results were achieved with α equals to 0.75 (i.e., the Euclidean distance should receive more importance than SDTW).

Comparisons. The best results of the proposed methods was achieved when γ is 0.05, $nVals$ is 85 and α is 0.75, achieving accuracy of 92.5%, 97.5% and 98.75% on lateral, oblique and frontal views, respectively. We believe the lower accuracy on lateral view is caused by occlusion, which is more common on that view. Euclidean distance method is more efficient than SDTW, showing that static information is more efficient than dynamic on gait recognition. However, the accuracy is increased when dynamic and static gait information from SDTW and Euclidean distance are fused.

The work is compared to others that use the same dataset. According to the results showed in Table 1, our approach has some of the best results on lateral, the second best on oblique and the best on frontal view (together with Kusakunniran et al. [9]). It is also interesting to note that our results are better than Liu et al. [11], which is a recent model-based method that uses deep learning. These results prove the efficiency of our method, despite the simplicity of the developed algorithms.

5 Conclusions

This work used pose estimation on gait recognition to retrieve body parts coordinates and transform them to signals and movements histograms, that are then used as features. Two methods were used for recognition. The first employs Subsequence Dynamic Time Warping to compare signals from the probe and gallery and rank them based on minimum matching cost; and the other uses Euclidean distance on the movement histograms to define the gallery sample closest to probe. A score fusion is then used on the methods results.

Experiments were performed on all views of CASIA Dataset A and we conclude that occlusion has a great effect on gait recognition. Therefore, noisy body parts should not be used on calculations. Euclidean distance was proved to be

better than Subsequence Dynamic Time Warping and the recognition is improved when the two methods are fused.

The accuracy of 92.5%, 97.5% and 98.75% was found for lateral, oblique and frontal views, respectively, which is compared with state-of-art works. This work is one of the best on lateral, the second on oblique, and the best on frontal (together with Kusakunniran et al. [9]), proving our work efficiency on gait recognition applications.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 311053/2016-5 and 438629/2018-3), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17), the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project) and Petrobras (Grant 2017/00643-0).

References

1. Bashir, K., Xiang, T., Gong, S.: Cross view gait recognition using correlation strength. In: Bmvc. pp. 1–11 (2010)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. vol. 1, p. 7 (2017)
3. Chen, C., Liang, J., Zhao, H., Hu, H., Tian, J.: Factorial hmm and parallel hmm for gait recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **39**(1), 114–123 (2009)
4. Derawi, M.O., Nickel, C., Bours, P., Busch, C.: Unobtrusive user-authentication on mobile phones using biometric gait recognition. In: Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on. pp. 306–311. IEEE (2010)
5. Hayfron-Acquah, J.B., Nixon, M.S., Carter, J.N.: Automatic gait recognition by symmetry analysis. Pattern Recognition Letters **24**(13), 2175–2183 (2003)
6. Isaac, E.R., Elias, S., Rajagopalan, S., Easwarakumar, K.: View-invariant gait recognition through genetic template segmentation. IEEE signal processing letters **24**(8), 1188–1192 (2017)
7. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology **14**(1), 4–20 (2004)
8. Kale, A., Cuntoor, N., Yegnanarayana, B., Rajagopalan, A., Chellappa, R.: Gait analysis for human identification. In: International Conference on Audio-and Video-Based Biometric Person Authentication. pp. 706–714. Springer (2003)
9. Kusakunniran, W., Wu, Q., Li, H., Zhang, J.: Automatic gait recognition using weighted binary pattern on video. In: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on. pp. 49–54. IEEE (2009)
10. Lee, H., Hong, S., Nizami, I.F., Kim, E.: A noise robust gait representation: motion energy image. International Journal of Control, Automation and Systems **7**(4), 638–643 (2009)

11. Liu, D., Ye, M., Li, X., Zhang, F., Lin, L.: Memory-based gait recognition. In: BMVC (2016)
12. Liu, J., Zheng, N.: Gait history image: a novel temporal template for gait recognition. In: Multimedia and Expo, 2007 IEEE International Conference on. pp. 663–666. IEEE (2007)
13. Man, J., Bhanu, B.: Individual recognition using gait energy image. IEEE transactions on pattern analysis and machine intelligence **28**(2), 316–322 (2006)
14. Mowbray, S.D., Nixon, M.S.: Automatic gait recognition via fourier descriptors of deformable objects. In: International Conference on Audio-and Video-Based Biometric Person Authentication. pp. 566–573. Springer (2003)
15. Nizami, I.F., Hong, S., Lee, H., Lee, B., Kim, E.: Automatic gait recognition based on probabilistic approach. International Journal of Imaging Systems and Technology **20**(4), 400–408 (2010)
16. Ogata, T., Tan, J.K., Ishikawa, S.: High-speed human motion recognition based on a motion history image and an eigenspace. IEICE TRANSACTIONS on Information and Systems **89**(1), 281–289 (2006)
17. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing **26**(1), 43–49 (1978)
18. Sarkar, S., Phillips, P.J., Liu, Z., Vega, I.R., Grother, P., Bowyer, K.W.: The humanid gait challenge problem: Data sets, performance, and analysis. IEEE transactions on pattern analysis and machine intelligence **27**(2), 162–177 (2005)
19. Wagg, D.K., Nixon, M.S.: On automated model-based extraction and analysis of gait. In: Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on. pp. 11–16. IEEE (2004)
20. Wang, L., Ning, H., Tan, T., Hu, W.: Fusion of static and dynamic body biometrics for gait recognition. IEEE Transactions on circuits and systems for video technology **14**(2), 149–158 (2004)
21. Wang, L., Tan, T., Hu, W., Ning, H.: Automatic gait recognition based on statistical shape analysis. IEEE transactions on image processing **12**(9), 1120–1131 (2003)
22. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE transactions on pattern analysis and machine intelligence **25**(12), 1505–1518 (2003)
23. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. vol. 4, pp. 441–444. IEEE (2006)
24. Zhang, E.H., Ma, H.B., Lu, J.W., Chen, Y.J.: Gait recognition using dynamic gait energy and pca+ lpp method. In: Machine Learning and Cybernetics, 2009 International Conference on. vol. 1, pp. 50–53. IEEE (2009)