# Pixel Volumes and Gait Recognition using Features of Balance between Model and Appearance

Wenqiang Liu [a], Tianqi Yang [b], Xiaofang Wu [c] and Zhongchao Xia [d]

School of Information Science and Technology, Jinan University, Guangzhou 510632, China;

[a]wenqiang0946@qq.com, [b]y_tq@163.com, [c]739327953@qq.com, [d]512801539@qq.com

## Abstract

**Walking at different speeds or carrying a backpack can change the walking posture of people, which is one of the important factors that restrict the performance of gait recognition methods. In response to this problem, we propose a new gait feature called pixel volume (PV), and use the support vector machines(SVMs) with dynamic time warping(DTW) kernel to identify. We used Kinect to record a dataset containing four variants of normal speed walking, fast walking, slow walking, and walking with a backpack. A part of the normal speed walking sequence was used as the galleries, the rest were used as probes, and assess the recognition rate in each case. The experimental results show that the proposed method not only achieves high performance under normal walking conditions, but also is robust in the case of changes in pace and carrying a backpack.**

## Keywords

**Gait recognition, pixel volume, dynamic time warping, support vector machine.**

## 1. Introduction

Gait recognition is intended to use people's walking posture for identification. Compared with biometrics such as fingerprint, iris and face, gait has some excellent characteristics. For example, gait recognition can be performed at a greater distance and does not require the deliberate cooperation of subjects. In addition, gait recognition is safer, and even if you have the same body shape as the one allowed, its dynamic walking posture can hardly be imitated. Therefore, gait recognition has great application potential in the fields of security and intelligent monitoring.

However, human gait can be disturbed by various factors, such as walking speed, carrying and clothing. Walking speed does not change the appearance of people, but it does have a big impact on the dynamic walking posture. Carrying and dressing mainly change the appearance of the person, and if the carry is heavy, it will also change the gait. The existing gait recognition methods can be roughly divided into two categories, appearance-baseds method and model-based methods. Appearance-based methods are sensitive to changes in clothing and carry-ons, but when pace changes or carrying heavy objects, they can also be distinguished by body shape information. The model-based approach is just the opposite.

In fact, most of the models are obtained through appearance, and there is no clear dividing line between them. For example, the elliptical model[1] is obtained by fitting the binary silhouette of different parts of the human body with ellipses. The model-based approach discards the details of the appearance after building the model. If the appearance-based approach preserves only the overall information of different areas of the human body, it can also be considered as a model-based approach.

In this article, we provide a flexible feature extraction method to achieve a balance between appearance and model. We use the depth image captured by Kinect to reconstruct the 3D point cloud on the front of the human body. Then calculate the volume occupied by each point in the point cloud, i.e. the pixel volumes. The human body area is divided into a plurality of blocks, and the sum of the pixel volumes in each block is used as the feature. The weights of the appearance and model is changed by adjusting the block size so that the information of the body shape is retained while ignoring the clothing. In addition, we also utilize the dynamic time warping[2] algorithm to minimize

gait differences due to speed changes and identify them with support vector machines. Finally, we collected a gait database containing normal speed walking, fast walking, slow walking, and walking with a backpack, and evaluated the recognition performance of the proposed method for the four variants.

## 2. Related works

The appearance-based approach uses the features (such as outlines, textures, etc.) of a sequence of walking images to represent the person's gait pattern. Gait energy image(GEI)[4] is one of the most successful appearance-based gait recognition methods. In [4], a sequence of standardized binary silhouette images are aligned and temporally averaged. Then GEI are encoded, dimensionally reduced, and identified. GEI is very successful for side gait images, but it would fail while perspective changes. Some methods[5] use multiple cameras to compensate for perspective defects, but the data collection cost is very high, and the subsequent data calculations are also very expensive. In [7], the author extended GEI to three-dimensional space and proposed a method for gait recognition using gait energy volume(GEV). They used depth images to reconstruct a partial volume of the front of the human body, then processed and identified it using the method like GEI. This method has good performance in frontal view and can adapt to some changes in perspective.

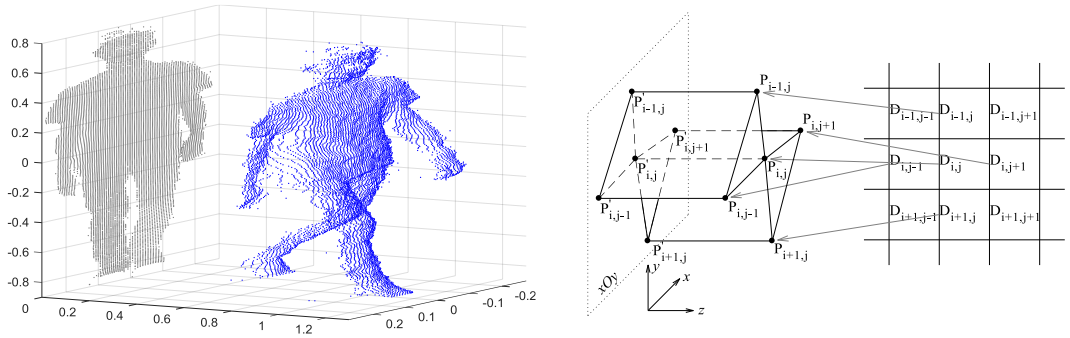
The model-based method characterizes the gait pattern by constructing a body motion structure model and mapping the gait image onto the structural components of the model. The Kienct skeleton model is one of the most representative models. It is obtained by mapping a body depth image to a human motion model consisting of 25 points. A demonstration of gait recognition using the Kinect skeleton model is given by Preis[8] et al. They use skeletal model data to calculate 11 static features such as height, leg length, and arm length, as well as 2 dynamic features of step size and speed, then for gait recognition. Although the performance is general, the feasibility is shown. In [9], the authors express human gait by calculate horizontal and vertical distances of selected joint pairs during one gait cycle and classify it with the K-nearest neighbor classifier. The method achieved 92% accuracy in a dataset with 20 subjects. However, the skeletal model is derived from depth images, and Kinect attempts to bring everyone's skeletal model closer to the built-in template, which severely diminishes the differences between individuals. Therefore, as the number of subjects increases, the classification ability of features obtained from the skeleton model will rapidly decreased.

Compared with the previous methods, we propose a flexible feature extraction method that can adjust the weights of appearance and model and try to balance them to achieve optimal gait recognition performance.

## 3. Proposed method

### 3.1 Pixel volumes

The mapping table provided by Kinect can be used to map the depth image into a spatial point cloud, where each point can be regarded as a volume voxel, i.e. the pixel volume. We define the pixel volume of a point as follows: Let the pixel of the *i*-th row and the *j*-th column in the depth image be $D_{i,j}$, the corresponding spatial point is $P_{i,j}$, and the projection point of $P_{i,j}$ on the xOy plane is $P'_{i,j}$, then the pixel volume is the sum of the volumes of the space bodies $P_{i,j}$-$P_{i-1,j}$-$P_{i,j-1}$-$P'_{i,j}$-$P'_{i-1,j}$-$P'_{i,j-1}$ and $P_{i,j}$-$P_{i+1,j}$-$P_{i,j+1}$-$P'_{i,j}$-$P'_{i+1,j}$-$P'_{i,j+1}$. As shown in Fig. 1.



(a) (b)

Fig. 1 (a) Spatial point cloud and its projection. (b) The pixel volume of $D_{i,j}$ is composed of two truncated triangular prisms: $P_{i,j}$-$P_{i-1,j}$-$P_{i,j-1}$-$P'_{i,j}$-$P'_{i-1,j}$-$P'_{i,j-1}$ and $P_{i,j}$-$P_{i+1,j}$-$P_{i,j+1}$-$P'_{i,j}$-$P'_{i+1,j}$-$P'_{i,j+1}$

The pixel volume of $D_{i,j}$ is calculated as:

$$\begin{aligned} V_{i,j} &= \frac{1}{2}\left|\overrightarrow{P'_{i,j}P'_{i-1,j}} \times \overrightarrow{P'_{i,j}P'_{i,j-1}}\right| \times \left[\left(\left|\overrightarrow{P_{i,j}P'_{i,j}}\right| + \left|\overrightarrow{P_{i-1,j}P'_{i-1,j}}\right| + \left|\overrightarrow{P_{i,j-1}P'_{i,j-1}}\right|\right)/3\right] \\ &+ \frac{1}{2}\left|\overrightarrow{P'_{i,j}P'_{i+1,j}} \times \overrightarrow{P'_{i,j}P'_{i,j+1}}\right| \times \left[\left(\left|\overrightarrow{P_{i,j}P'_{i,j}}\right| + \left|\overrightarrow{P_{i+1,j}P'_{i+1,j}}\right| + \left|\overrightarrow{P_{i,j+1}P'_{i,j+1}}\right|\right)/3\right] \end{aligned} \tag{1}$$

To eliminate the influence of the overall position of the body on the pixel volume, we align the barycentric of the point cloud with the origin by coordinate transformation. The barycentric coordinate calculation formula is:

$$P_0(x_0, y_0, z_0) = \left(\frac{1}{n}\sum_{k=1}^{n} x(k), \frac{1}{n}\sum_{k=1}^{n} y(k), \frac{1}{n}\sum_{k=1}^{n} z(k)\right) \tag{2}$$

Where n is the number of points and the coordinates after translation are:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -x_0 \\ 0 & 1 & 0 & -y_0 \\ 0 & 0 & 1 & -z_0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \\ 1 \end{bmatrix} \tag{3}$$

Then the aligned pixel volume formula is:

$$V'_{i,j} = V_{i,j} - \frac{1}{2}\left(\left|\overrightarrow{P'_{i,j}P'_{i-1,j}} \times \overrightarrow{P'_{i,j}P'_{i,j-1}}\right| + \left|\overrightarrow{P'_{i,j}P'_{i+1,j}} \times \overrightarrow{P'_{i,j}P'_{i,j+1}}\right|\right) \times z_0 \tag{4}$$

If one or more points constituting the truncated triangular prism do not belong to the body or do not exist, the volume of this portion is zero. Fig. 2 illustrates a schematic representation of the body pixel depth volume projected on the xOy plane.

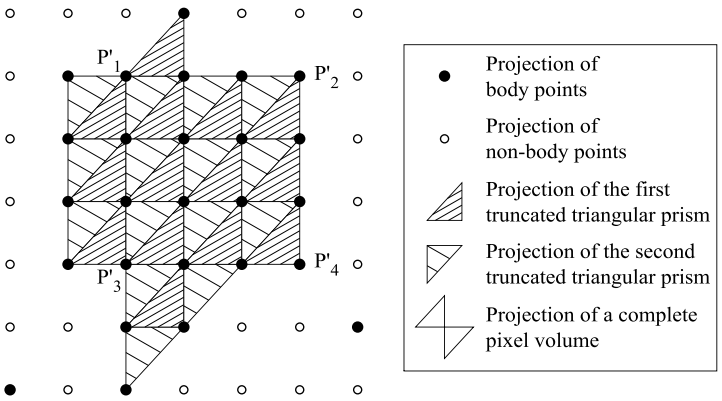



Fig. 2 $P'_1$ has only the volume of the second part, $P'_2$ has a volume of 0, $P'_3$ has a volume of both two parts, and $P'_4$ has only the volume of the first part

### 3.2 Feature extraction

To embody the motion patterns of different parts of the human body, we divide the body region into several blocks, and use the sum of the pixel volumes of all points in the block as gait features.The projection of the human body region in xOy plane is divided into blocks of M rows and N columns, and the height and width of the blocks are $l_1$ and $l_2$. The block located in the *m*-th row and the *n*-th column is denoted as B(m,n). If the projection of the point $P'_{i,j}$ is within B(m,n), then $P'_{i,j} \in$ B(m,n). Fig. 3 explains the method of block partitioning.

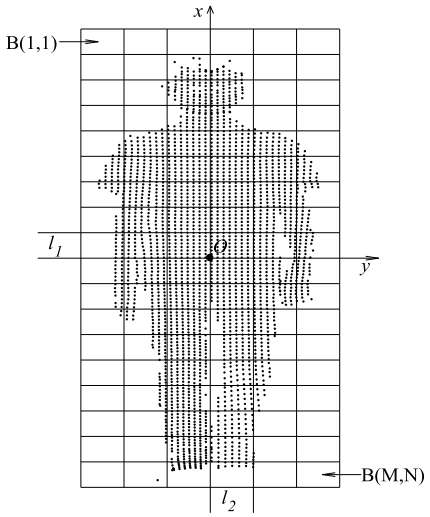



Fig. 3 Block partitioning

For each block B(m,n), its gait feature is:

$$V_{B(m,n)} = \sum_{P_{i,j} \in B(m,n)} V'_{i,j} \tag{5}$$

For each frame of data, its gait feature vector is:

$$v = \left[ V_{B(1,1)}, V_{B(1,2)}, \cdots, V_{B(m,n)}, \cdots, V_{B(M,N-1)}, V_{B(M,N)} \right]^T \tag{6}$$

### 3.3 Classification

The speed of walking and the change in step size make the length of gait cycle sequences vary widely. Traditional support vector machines can only classify feature vectors with the same dimensions. In[10], Bahlmann et al. replaced the radial basis kernel with the GDTW kernel, enabling support vector machines to efficiently classify feature vectors in different dimensions of handwritings. This can also be used in gait recognition.

The dynamic time warping algorithm measures the similarity between two gait feature vector sequences $T = (t_1, t_2, \cdots, t_{L_T})$ and $R = (r_1, r_2, \cdots, r_{L_R})$ by using dynamic time warping distance. Given a warping path $\phi = (\phi(1), \cdots, \phi(i), \cdots, \phi(n))$ (s.t. $\phi(i) = (\phi_T(i), \phi_R(i)) \in \{1, \cdots, L_T\} \times \{1, \cdots, L_R\}$ , $\phi_T(i) \le \phi_T(i+1) \le \phi_T(i)+1$ , $\phi_R(i) \le \phi_R(i+1) \le \phi_R(i)+1$) and a distance formula d, such as $d(t_i, r_j) = \|t_i - r_j\|^2$ , the warping path distance is:

$$D_\phi(T,R) = \frac{1}{n} \sum_{i=1}^{n} d\left(t_{\phi_T(i)}, r_{\phi_R(i)}\right) \tag{7}$$

The goal of dynamic time warping is to find a warping path $\phi^*$ that minimizes the distance of the regular path, i.e.

$$D(T,R) = D_{\phi^*}(T,R) = \min_{\phi} \{ D_\phi(T,R) \} \tag{8}$$

The GDTW kernel function replaces the $\|x - x'\|^2$ in the radial basis kernel function with the dynamic time warping distance, i.e.

$$K(x,x') = \exp\left(-D(x,x')/2\sigma^2\right) \tag{9}$$

## 4. Experiments and results

### 4.1 Dataset

We collected a gait dataset with 152 subjects, each containing 12 walking sequences. For each subject, we set four variants of normal speed walking (N1-N6), faster walking (F1-F2), slower speed walking (S1-S2) and carrying a 3kg backpack walking (B1-B2). The experiment uses gait cycle sequences separated from N1-N4 as galleries, and the rest as probes, and evaluated the recognition performance for the four variants.

### 4.2 Parameter settings and Results

In KNN, k=1. We divided the 2.00m×0.80m body regions with blocks of 0.02m×0.02m(small), 0.05m×0.05m(medium) and 0.10m×0.10m(large) and evaluated the recognition rates under these three sizes. The experimental results are shown in Fig. 4.

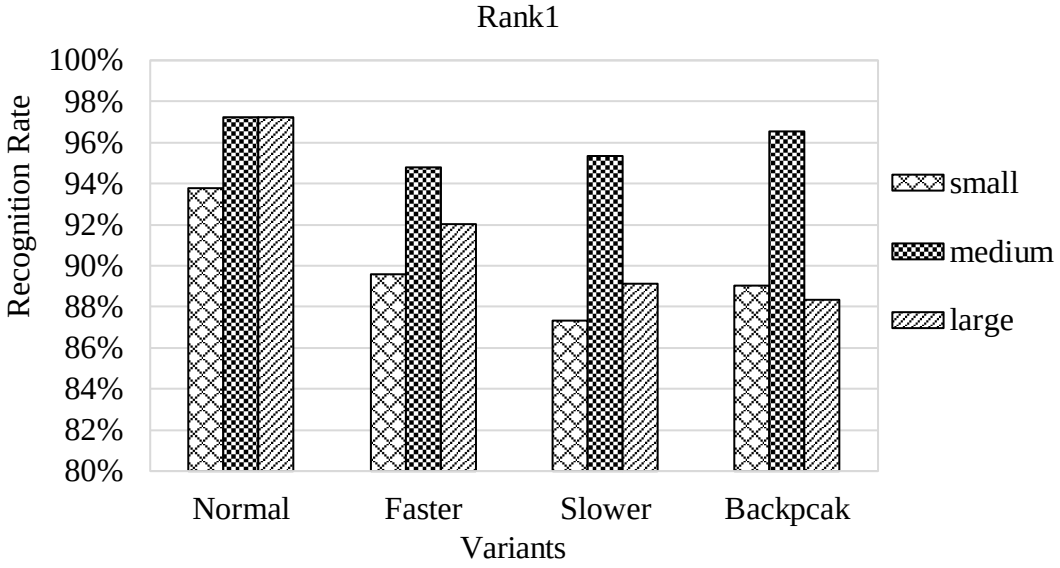



Fig. 4 Recognition rate of different block sizes in four variants.

It can be seen that "medium" has the highest recognition rate among all variants. The experiment achieved a high recognition rate in the normal condition, but in faster, slower and backpack, the recognition rate under every block size was significantly reduced except for "medium". It is worth noting that "medium" and "large" have the same recognition rate in normal, but in other cases, "large" is lower than "medium". This is because “large” highlights the model information, while different speeds and backpacks just changed them. In "small", the recognition rate is not good in all cases, which indicates that over-emphasizing the appearance details is not conducive to the improvement of recognition performance.

### 4.3 Compared with other methods

For comparison, we performed the same experiments on the datasets we collected using the GEI, GEV, and HDF+VDF methods. The experimental results are shown in Table 1.

Table 1 Comparison of recognition rates of different methods.

| Method | Normal | Faster | Slower | Backpack | Total |
|---|---|---|---|---|---|
| GEI | 82.03 | 69.02 | 74.32 | 73.06 | 74.91 |
| GEV | 91.01 | 81.6 | 82.30 | 79.31 | 83.54 |
| VDF+HDF | 25.35 | 12.27 | 18.09 | 15.95 | 18.24 |
| Our Method | 97.24 | 94.79 | 95.33 | 96.55 | 96.03 |

* Our method uses the block size of 0.05m×0.05m.

Both the GEI and GEV methods performed well under normal conditions, but in the case of faster, slower, and backpack, the recognition rates of both methods dropped by about 10%. The GEVs performs better than GEIs because the frontal depth image contains more 3D information. The

VDF+HDF method performed poorly, while in [9] they reported a 92% recognition rate for a dataset of 20 subjects, which also validated the aforementioned related inferences.

## 5. Conclusion

In this paper, we propose a new gait feature called the pixel volumes witch balances model information and appearance information. The human body area is divided into blocks of appropriate size, and the sum of pixel volumes in each block is taken as the final gait feature. Finally, the support vector machine with dynamic time warping kernel is used for classification and recognition. Our method achieved 97.24%, 94.79%, 95.33% and 96.55% recognition rates in walking sequences of normal speed, faster speed, slower speed, and with a 3kg backpack, respectively. This shows that the proposed method is robust to pace changes and carrying items. In addition, we also compared our method with the GEI, GEV and VDF+HDF methods, and the results show that the proposed method obtained a better recognition performance than the comparison methods.

## Acknowledgements

This work was supported by the Science and Technology Project of Guangdong, China (Grant No.2017A010101036).

## References

[1] Lee L, Grimson WEL: Gait Analysis for Recognition and Classification, *Proc. of the IEEE International Conf. on Automatic Face Gesture Recognition*(Washington, UAS, May. 21-21, 2002). p.155-162.

[2] Sakoe H, Chiba S: Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics Speech & Signal Processing, Vol.26 (2003) No.1, p.43-49.

[3] Paliwal KK, Agarwal A, Sinha SS: A Modification over Sakoe and Chiba's Dynamic Time Warping Algorithm for Isolated Word Recognition, Signal Processing, Vol.4 (1982) No.4, p.329-333.

[4] Han J, Bhanu B: Individual Recognition using Gait Energy Image, IEEE Transactions on Pattern Analysis & Machine Intelligence, (2006) No.2, p.316-322.

[5] Shakhnarovich G, Lee L, Darrell T: Integrated Face and Gait Recognition from Multiple Views, *Proc.of the IEEE Computer Society Conf. on Computer Vision & Pattern Recognition*(Kauai, Hawaii, Dec. 8-14, 2001). Vol.1, p.I-439-I-446.

[6] Gu J, Ding X, Wang S, et al: Action and Gait Recognition from Recovered 3-D Human Joints, IEEE Transactions on Systems Man & Cybernetics Part B, Vol.40 (2010) No.4, p.1021-1033.

[7] Sivapalan S, Chen D, Denman S, et al: Gait Energy Volumes and Frontal Gait Recognition using Depth Images, *International Joint Conference on Biometrics*(Washington, USA, Oct. 11-13, 2011). p.1-6.

[8] Preis J, Kessel M, Werner M, et al: Gait Recognition with Kinect, *1st international workshop on kinect in pervasive computing*. (Newcastle, UK, Jun. 18-18, 2012). p.1-4.

[9] Ahmed M , Al-Jawad N , Sabir A T: Gait Recognition Based on Kinect Sensor, *Real-Time Image and Video Processing 2014*(Brussels, Belgium, Apr. 16-17, 2014). Vol.9139, p91390B.

[10] Bahlmann C, Haasdonk B, Burkhardt H: On-Line Handwriting Recognition with Support Vector Machines - A Kernel Approach, *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*( Ontario, Canada, Aug. 6-8,2002). p.49-54.