Does My Gait Look Nice? Human Perception-based Gait Relative Attributes Estimation by Dense Trajectory Analysis

Allam Shehata, Yuta Hayashi, Yasushi Makihara, Daigo Muramatsu, and Yasushi Yagi

Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
{allam,hayashi,makihara,muramatsu,yagi}@am.sanken.osaka-u.ac.jp
allam@eri.sci.eg

Abstract. Relative attributes have achieved remarkable contributions to object recognition and image classification tasks. These attributes provide high-level semantic explanations to describe and relatively relate the objects to each other instead of using direct labels to each object. In this paper, we take the first step in exploiting the gait relative attributes estimation. We propose firstly to build a robust gait motion representation based on the extracted dense trajectories (DTs) from gait videos, which is more suitable for gait attribute estimation than the existing heavily body shape-dependent appearance-based features such as gait energy image (GEI). More specifically, we apply a Fisher vector (FV) encoding framework to histogram of optical flows (HOFs) computed along with individual DTs. Afterward, we create a novel gait data set which contains 1,200 walking subjects and annotation of gait relative attributes based on human perception for gait pairs from the subjects. To estimate the relative attribute, we train a set of ranking functions for the relative attributes using Rank-SVM classifier. These ranking functions estimate a score that implies the strength of the presence of the attributes for each walking subject. The experimental results showed that the proposed method could well represent gait attributes and also that the proposed gait motion descriptor had better generalization capability than GEI for gait attributes estimation task.

Keywords: Dense trajectories · Relative attribute · Gait attribute estimation · Ranking functions · Histogram of optical flow · Fisher vector.

1 Introduction

Nowadays, the walking style, i.e. gait, of the pedestrian has been adopted as a powerful biometric modality. This is because the gait can be recorded remotely without human participation or any need for special sensors. Several gait-based approaches have been proposed recently to serve many applications [20,28]. These applications include but not limited to; gait analysis/recognition [33], age/gender estimation [19,21], person identification, and person verification [18,20]. The existing human gait recognition systems adopt two main types of modeling: model-based and model-free. The model-based approaches typically model the kinematics of human joints to measure physical gait parameters such as step lengths, and angular speeds [2]. As a consequence,

model-based models suffer from high computational costs due to parameters calculation. They also influence by the quality of gait sequences. In the model-free category, a compacted representation for the gait sequences is considered without explicit modeling of body structure [24]. These models use the features extracted from the motion or shape of the walking subject and hence require much less computation. They are, however, not always robust against covariates such as viewpoints, clothing, carrying status, and so on [29]. Both of the aforementioned modeling types require some preprocessing steps to extract binary silhouettes, skeletons, or body joints to encode the gait motion/appearance information. This limits the performance of the approaches due to the presence of dynamic backgrounds, noisy segmentation, and inaccurate body joints localization.

Instead, gait motion information can be computed directly using the optical flow through spatial-temporal domain [36]. Based on this, the local motion descriptors are proposed and have become popular in human action recognition community [4]. To build these descriptors, highly dense sample points are detected and tracked through the action video. Afterward, then motion information relevant to these points are aggregated in the form of histograms. For instance, the recent work that introduced in [4] used the local motion descriptors based on dense trajectories for gait recognition instead of binary silhouettes. They propose to extract the local motion features on different body regions and combine the extracted descriptors into a single global descriptor using the Fisher vector encoding mechanism [26].

Although the satisfactory performance that achieved by the gait analysis/recognition approaches, to the best of our knowledge, the problem of gait attributes estimation based on human perception and relating the persons to each other based on their gait attributes have not yet widely explored. An early attempt is introduced in [33] to figure out the encoding process of the socially relevant information (i.e. human attributes) into biomedical motion patterns. They proposed to introduce an interactive linear classifier tool that can discriminate male from female walking patterns given the instantaneously changes in a set of defined human attributes such as age, gender, and mode. Furthermore, another method is proposed in [11] which adopts the attribute-based classification to overcome the computational complexity of the traditional multi-class gait classification task. This work mainly tries to define some attributes based on the similarities between subjects classes. It is, however, applied only for gait recognition improvement and does not handle the task of gait attributes estimation. Another recent work introduced in [39] proposes to use convolutional neural networks and multi-task learning model to identify the human gait and simultaneously estimate his/her attributes. The authors claim that their work is the first multi-attributes gait identification system being proposed. Nevertheless, in their representation, they still depend on the GEI-based representation which suffers from their special considerations and the body shape changes influences.

Currently, with deep learning models, it became easier to identify the person based on his/her gait style once you have sufficient training of walking sequence [**?**,**?**]. It will be, however, hard for such models to automatically recognize and identify the person's gait, when he/she has never been seen before (i.e., in a zero-shot learning scenario). Moreover, several classification-based systems adopt directly associating low-level features with absolute annotation labels. By contrast, discriminative visual descriptions

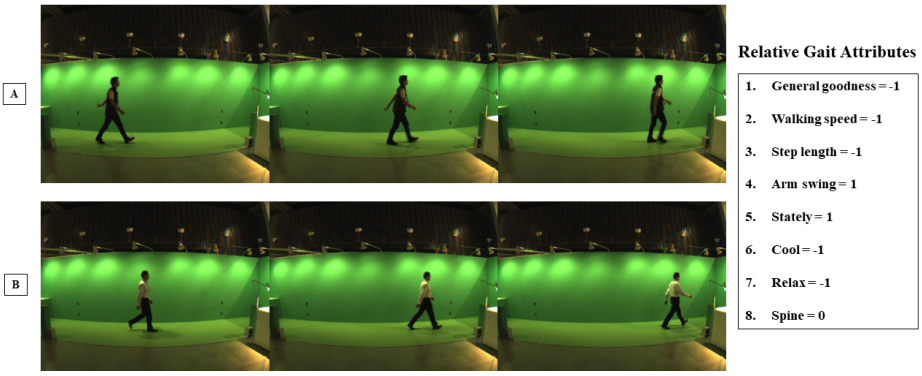

Fig. 1: The gait relative attributes for two walking subjects A and B. Eight relative attributes are considered. 1 means that the attribute strength in person A is greater than in B. -1 means attribute strength in person B is greater than in A. 0 means that both attribute strengths in A and B are almost the same.

cannot be characterized easily by an absolute label. Therefore, to describe action properties, it would be better to define high-level semantic concepts in terms of relative attribute labels instead of absolute ones. Recent works prove that these attributes can be adopted in human recognition tasks as an intermediate level of descriptions for the human visual properties [40]. Following that, Parikh et al. [25] proposed the concept of relative attributes to propose the learning to rank classification model. They use the associated strengths of relative visual features in objects/scenes to learn a set of ranking functions and use these functions to describe/relate new unseen instances to the training images. the proposed method is examined on different images datasets of faces and natural scenes achieving an excellent performance over traditional binary attribute prediction methods.

In our proposed work, we exploit the concept of the relative attributes to build the first human gait relative attributes estimation framework. This framework aims at assessing the gait styles of walking subjects pairs based on a set of gait relative attributes as shown in Fig 1. Given the attributes annotation labels of the training pairs, we learn a set of ranking functions based on Rank-SVM classifier [25] and use these ranking functions to relate the subjects to each other and predict the gait relative attributes for the unseen subjects. For gait motion encoding, we use the Fisher vector encoding scheme to build a robust gait motion representation based on the extracted dense trajectories from the gait videos.

The main contributions of this work can be summarized as follows:

1. **Introduce the first gait relative attributes estimation approach**
   In gait-based approaches, the concept of relative attributes has never been explored yet. We propose to exploit this concept and introduce an approach to estimate the gait relative attributes of the walking subjects based on human perception.
2. **Build a new gait data set with attributes annotation**

We introduce a new annotated gait dataset based on human perception. We hire seven annotators to watch the walking videos for 1,200 walking subjects. We defined eight gait attributes and asked the annotators to assess each attribute for each pair of subjects based on the annotator human perception. This evaluation appears in the form of annotation labels that express the presence of an attribute for each subject pair (see Fig 1). We use this annotation to train and evaluate our proposed gait relative attributes estimation methods.

3. **Propose a robust dense trajectory-based gait motion representation**
   Instead of using the existing and widely used appearance-based gait descriptors such as GEI [9], we exploit the dense trajectories to encode the motion information of the gait dynamics. We build a histogram of optical flow (HOF) for each trajectory. Then, these HOFs are encoded into a single robust global motion descriptor using the Fisher vector encoding. This global vector carries the gait motion of the entire walking sequence.

The rest of the paper is organized as follows. Section 2 outlines the most related work. In Section 3, a detailed explanation of our proposed motion feature representation. The ranking learning process, the relative attributes annotations,an the proposed evaluation framework are provided in Section 4. The detailed experiments are included in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

Currently, the gait recognition systems can be classified into two main types: model-based and model-free approaches. In model-based approaches, a predefined model is introduced to model the human bodies' movement [3]. The model-free approaches depend on the image features directly (i.e., appearance) for analysis and recognition. The GEI descriptor has been adopted by most of the gait recognition approaches as a baseline to represent the person's walking style. GEI extracted by averaging over the binary silhouette sequence of the walking person [9]. It mainly depends on extracting pure foreground pixels and its robustness may be degraded at severe occlusion, moving camera, and scene clutter conditions.

To partially overcome such aforementioned limitations, new descriptors based on the local motion feature of sampling points instead of binary silhouettes were considered. These descriptors mainly embraced in the field of human action recognition [36,5]. A new approach is proposed in [36] to describe videos by dense trajectories (DTs)[1] by sampling dense points from each frame and track them based on displacement information from a dense optical flow field. The resulted trajectories are robust to fast irregular motions and cover the motion information in videos properly. Moreover, a new motion descriptor is designed based on differential motion scalar quantities, divergence, curl and shear features (DCS). This descriptor adequately decomposing visual motion into dominant and residual motions, both in the extraction of the space-time trajectories.

[1] Through the manuscript, we will refer to the dense trajectory by DT, gait energy image by GEI, histogram of optical flow by HOF, and Fisher vector by FV.

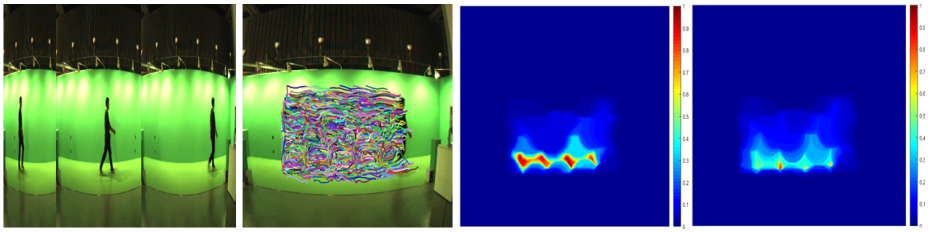

Fig. 2: (a) Sampled frames from the walking sequence, (b) Collected DTs from the video sequence, (c) location probability map for trajectories distribution, and (D) Motion magnitude probability map.

Recently, local feature descriptors are leveraged to serve in gait-based analysis/ recognition system. The space-time interest points (STIPs) [14] are used for representing the person's gait in [12]. On the other hand, the gait modality is utilized for enhancing the multi-person tracking through crowded scenes based on a trajectory-based clustering [32]. The key novelty of this method is adopting the person's individuality, that is, the gait features (in the frequency domain) and the temporal consistency of local appearance to track each individual in a crowd. The experiments show that the usage of these combined features contributes to significant performance improvement in tracking people in crowded scenes. A new motion descriptors is proposed in [5,17,23] for improving gait recognition based on DTs-based pyramidal representation. The person region is divided into sub-regions and the local features are extracted and then combined into a single high-level gait descriptor by using the Fisher Vector encoding [27]. Recently attribute-based methods prove the usefulness of object attributes as high-level semantic description. It serves well in the action recognition and zero-shot learning [16,25,1]. A method proposed in [7] to shift the goal of object recognition to the description by describing the objects and learn from these descriptions. They claim that if the object's attributes are adopted as the anchor representation for object recognition systems, so more details about the objects can be revealed than just its name. The direct attribute prediction (DAP) model is introduced in [13] to predict the presence of each attribute to train object models. The relative attribute concept is proposed in [25] to semantically describing the relationships between the objects based on a set of defined attributes.

3 The Proposed Approach

In the following subsections, we explain the proposed DTs-based gait motion representation in details. As well, we describe the learn to rank model based based on the proposed gait relative attributes annotations.

3.1 Dense Trajectories Extraction from Gait Videos

Recently, the local motion feature descriptors have been utilized in the field of action recognition [36,22]. These descriptors basically are built on the extracted short fragment trajectories of the tracked sampled points. They often can be extracted directly

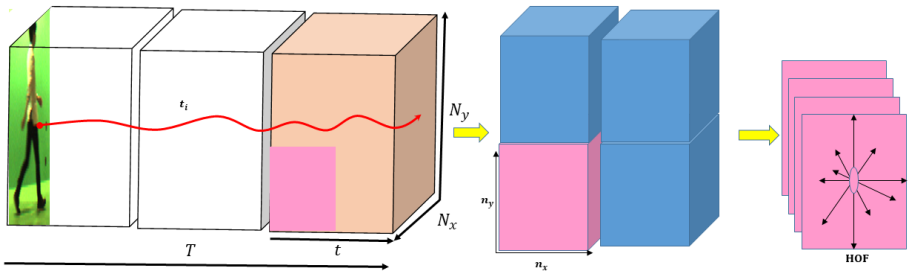

Fig. 3: HOF descriptor computation over $N_x \times N_y \times T$ spatial-temporal volume along trajectory t_i. $N_x \times N_y$ is local neighborhood pixels and T is the temporal length. The volume divided into 3 main blocks ($N_x \times N_y \times t$). To capture the actual motion dynamics, every block is subdivided into four ($n_x \times n_y \times t$) sub-blocks. For each sub-block, the HOF descriptor is computed. The global HOF of the trajectory is the concatenation of the resulted HOFs from sub-blocks.

from the original frames without pure foreground segmentation as shown in Fig. 2(b). These trajectories best describe the actual spatial-temporal dynamics of the moving objects and emerge their kinematics properly (see Fig. 2(c,d)). Currently, these local feature descriptors are utilized in gait recognition systems [5,23,37] and achieved remarkable enhancements to the gait analysis/recognition tasks. In our work, we extract a set of dense trajectories (DTs) from the walking videos using the state of the art DTs extractor [36]. We use the extracted DTs to build a global gait motion descriptor for the gait styles of the walking subjects.

3.2 Dense Trajectory-based HOF Descriptor

HOF descriptor has become widely used in many recognition and video classification approaches [34,35,36,15]. To encode the the motion information of dense trajectories, we propose to compute the HOF descriptors from the 3D spatial-temporal volume around the DTs. As shown in Fig. 3, we assign $N_x \times N_y \times T$ volume along the trajectory where $N_x \times N_y$ is the spatial dimension and T is temporal length the of the DT. We extract $N_x \times N_y$ spatial window around each xy point lies on the DT considering this point the window anchor. These extracted windows are stacked to form the 3D volume. Notice that, these extraction criteria enable us to embed the structure information of the DT properly. Afterward, the optical flow displacement vectors in horizontal and vertical directions are computed for each 3D volume using the classical Farneback optical flow method [8]. As well, the average velocity is computed for each trajectory and subtracted from displacement vectors to suppress the fast irregular motions. Both of the average speed-subtracted displacement vectors are 2-dimensional vector fields per frame, so we can use them to compute the corresponding motion magnitudes and orientations. For robust motion encoding, we divide the 3D volume into 3 main blocks ($N_x \times N_y \times t$) and every main block is in turn again subdivided into four ($n_x \times n_y \times t$) sub-blocks. Note that, $t = T/3$ and $n_x = N_x/2$ and these values are experimentally selected for best performance. Now the number of sub-blocks in the entire 3D volume

is 12 sub-block[2], for each sub-block we compute a HOF Histogram. Given the obtained magnitudes and orientations grids, the magnitude is quantized in 9-bin orientations using full orientations and aggregated over every sub-block in both spatial and temporal directions as shown in Fig. 3. We concatenate all of adjacent sub-blocks responses into one global HOF descriptor. We normalize this global descriptor using its **L1**-Norm followed by signed square root [35]. For now, each DT is represented by its corresponding global 108-dimensional HOF descriptor. As a result, the motion information of the entire walking video is encoded into a group of global HOFs descriptors.

3.3 Fisher Vector Encoding for Gait Motion

The Fisher vector encoding (FV) mechanism popularly used in visual classification tasks to represent the image/scene as a set of pooled local descriptors. Principally, it follows the fisher kernel principle of estimating the class for a new object by minimizing the average of the Fisher kernel distance between known and unknown classes [10]. FV encode the local D-dimensional descriptors based on a trained generative model, which is often the Gaussian mixture model (GMM).

Let $H = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N)$ be a set of D-dimensional HOFs descriptors extracted from the walking video, one HOF for each DT. Suppose that the K-mixtures GMM is considered to model the generative process of these local descriptors. So, we estimate the GMM parameters by fitting it to H descriptors. Given this K-mixtures GMM, where each mixture has its mixture weight, mean, and covariance parameters respectively, $(\pi_k, \boldsymbol{\mu}_k, \Sigma_k)$. Let Θ denotes the vector of Q parameters of the K-mixtures GMM, $[\theta_1, \ldots, \theta_Q]^T \in \mathbb{R}^Q$ where $Q = (2D + 1)K - 1$. Given each θ expressed as $(\pi_k, \boldsymbol{\mu}_k, \Sigma_k)$, so the parameter vector Θ can be represented as follows

$$\begin{aligned} \Theta &= (\pi_1, \ldots, \pi_K; \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K; \Sigma_1, \ldots, \Sigma_K) \\ &= (\theta_1, \ldots, \theta_Q). \end{aligned} \tag{1}$$

Following [26], the gradient vector of D-dimensional the HOFs descriptors w.r.t. the parameters of GMM can be written as

$$\begin{aligned} &\nabla_\theta \log f_\theta(H|\Theta) \\ &\qquad = \left(\frac{\partial \log f(H|\Theta)}{\partial \theta_1}, \ldots, \frac{\partial \log f(H|\Theta)}{\partial \theta_Q} \right)^T. \end{aligned} \tag{2}$$

Note that f is the probability density function of GMM which models the generative process of H and it is expressed as follows

$$f(\boldsymbol{h}|\Theta) = \sum_{k=1}^{K} \pi_k \exp\left(-\frac{1}{2} (\boldsymbol{h} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{h} - \boldsymbol{\mu}_k) \right). \tag{3}$$

The function $\log f(H|\Theta)$ in Eq. (2) is the log-likelihood of the density function of GMM and it can be rewritten as $\log f(H|\Theta) = \sum_{n=1}^{N} f(\boldsymbol{h}_n|\Theta)$. Roughly speaking, FV

[2] 3 main blocks and each one subdivided into 4 sub-blocks. Each sub-block has dimension $8 \times 8 \times 5$. HOF descriptor for each trajectory has a dimension $12 \times 9 = 108$.

is the concatenation of the gradients results from Eq. (2) for $k = 1, \ldots, K$. Finally, the FV is normalized by using power normalization criteria [26]. For the detailed derivation of Fisher vector encoding, we may refer the readers to [26]. In our work, we leverage this method to encode the gait motion information from the walking video into a single global descriptor.

4 Learning Gait Relative Attributes

In object recognition field, the visual properties of objects have become attractive. They provide a high-level semantic description of the objects and their explicit and implicit features. To learn such attributes from the given training sample, the researchers go in two directions. The first one is to learn the mapping of the examples to an ordered set of numerical ranks [30]. The other direction is to learn set ranking functions from relative ranking preferences between example pairs [25]. In our proposed work we adopt the second direction. Given the computed global FVs descriptors from the training gait videos, we seek to learn a set of ranking functions from these descriptors and then use these ranking functions scores to predict the relative attributes of unseen test samples. To learn such ranking functions, we use our novel human perception-based pairwise gait relative attributes annotations to build the ordered and un-ordered pairs matrices. We explain the proposed annotation in details in 4.1.

Given the walking subjects pairs $(A, B) = \{(a_1, b_1), \ldots, (a_z, b_z)\}$, $z = 1, \ldots, M$ and its corresponding relative attributes annotations. For each annotated attribute l_p, we select all the subjects pairs that have $l_p = 1$ or -1 and adopt them as ordered pairs. Then, we build the ordered pairs sparse matrix O that has the dimension $I \times n$, where I is the total number of ordered pairs and n is the number of subjects. This matrix has a row for each subject pair and each row only contains single 1 and single -1. For instance, if l_p strength in subject a is greater than in b (i.e., $a \succ b$), then the row i contains $O(i, a) = 1$ and $O(i, b) = -1$ respectively. Similarly, we select all the subjects pairs that have $l_p = 0$ To build the un-ordered pairs sparse matrix $\{U_p\}$ that as the dimension $J \times n$. If both subjects a and b has the similar attribute strength, then $U(j, a) = 1$ and $U(j, b) = -1$. Following the ranking process in [25], we denote a set of ranking functions to be learned for each of the annotated attributes. This ranking function can be defined as $r_p(h_i) = \boldsymbol{w}_p^T \boldsymbol{h}_i$ under satisfaction of the following constraints:

$$\begin{aligned} \forall (a, b) \in O_p : \boldsymbol{w}_p^T \boldsymbol{h}_a > \boldsymbol{w}_p^T \boldsymbol{h}_b \\ \forall (a, b) \in U_p : \boldsymbol{w}_p^T \boldsymbol{h}_a = \boldsymbol{w}_p^T \boldsymbol{h}_b. \end{aligned} \tag{4}$$

The ranking function is usually assumed to be a linear function [6]. Because of that, it appears in the form $\boldsymbol{w}_p^T \boldsymbol{h}_i$ where $\boldsymbol{w}_p$ is the zero-shot weights produced from the Rank-SVM classifier [25].

4.1 Gait Attributes Annotation

To prepare the annotations of gait relative attributes, we select 1,200 subjects' walking sequences from the publicly available OULP-Age dataset [38]. These sequences, then,

Table 1: Sample from our proposed gait relative attributes annotations based on the human perception of the annotators. $a \succ b$ means that strength of an attribute in subject a is greater than in subject b, and $a \prec b$ is the vice versa. However, $a \sim b$ means that both of subjects have the same attribute strength. 1 informs that $a \succ b$, -1 informs that $a \prec b$, and 0 informs that $a \sim b$.

Subjects Pairs	Annotators	General goodness	Stately	Cool	Relax	Arm swing	Step length	Walking speed	Spine
$28 \Leftrightarrow 30$	Annotator X	-1	1	-1	-1	1	-1	-1	0
	Annotator Y	-1	-1	-1	1	1	0	0	0
	Annotator Z	-1	-1	0	0	-1	-1	0	0
$1423 \Leftrightarrow 1479$	Annotator X	-1	-1	0	1	-1	0	-1	-1
	Annotator Y	1	1	0	0	-1	1	-1	0
	Annotator Z	0	-1	1	0	-1	0	0	-1

delivered to 7 annotators. Each annotator was asked to watch the binary silhouette sequences for 1,200 subjects pairs and report his/her observations for each pair based on human perception. We define eight gait attributes under the names {*General goodness, Stately, Cool, Relax, Arm swing, Walking speed, Step length, Spine*}. Each of these attributes explain a visual property of the walking subject and it can receive three different labels, $\{1, 0, -1\}$. The annotator should assign 1 to the attribute if he/she observed that strength of the attribute in gait style of subject a is greater than in subject b. Similarly, the annotator should assign -1 if b has attribute strength greater than a. If the annotator guessed that both subjects have the same strength of the attribute, he /she assigns the label 0. In Table 1, we report the relative attributes annotations for two subjects pairs for illustration. For each pair, the relative attribute labels are listed for three different annotators. One can easily observe that for the same gait attribute, different annotators assign different/same labels according to their human perception. For instance, for subject pair (28,30), Annotators X and Y observed that *Arm swing* in 28 is better than in 30 (i.e. assign 1). However, the annotator Z observed the opposite (i.e. assign -1). As well, all the annotators agree that both 28 and 30 have the same *Spine* (i.e. assign 0).

4.2 Evaluation Criteria

To evaluate the predicted attribute labels against the ground truth annotation, we derive a new multi-thresholds evaluation criteria. Notices that the learned ranking functions produces real valued ranks. We use these ranks to evaluate the decision of classifier. Given test descriptor $\boldsymbol{h}_z$ and the weights vector $\boldsymbol{w} \in \mathbb{R}^D$, which obtained from R-SVM training stage, we compute the predicted scores through the inner-product, $\boldsymbol{w}_p^T \boldsymbol{h}_z$. Suppose having a set of s thresholds $L = (L_1, \ldots, L_s)$ and $L_1 < L_2 < \cdots < L_s$. The real valued ranks has one score for each subject (i.e., $\boldsymbol{h}_z$). We suggest to compute the difference score for each pair and then threshold them to mapping to predicted attributes labels. For instance, the score difference between the ranking functions scores of subjects a and b is $d_{ab} = r_p(\boldsymbol{h}_a) - r_p(\boldsymbol{h}_b)$. There are three label need to be predicted

$(1, 0, -1)$, therefore, we derive the following multi-thresholds evaluation criteria

$$l_p = \begin{cases} 1 & \forall \quad d_{ab} > L_i, \quad i = 1, \ldots, s \\ 0 & \forall \quad d_{ab} \quad \nexists \quad d_{ab} > L_i \quad \wedge \quad d_{ab} < -L_i \\ -1 & \forall \quad d_{ab} < -L_i \end{cases} \tag{5}$$

In Eq. (5), the first condition used for the attribute assignment 1, the second condition for 0 and the third for -1. For instance, the attribute l_p for a specific subject pair will be 1, if the score difference for subjects pairs (a, b) is greater than L_i and hence, we predict that $a \succ b$ for that attribute. Given the computed scores differences, we map them to attribute assignments using Eq. (5) and counts the mapped assignments for each threshold. To evaluate the predicted attributes against the GT annotation, we compute the accuracy precision measure for each threshold and adopt the highest precision.

5 Experiments

We evaluate our approach on a selective gait dataset from the publicly available OULP-Age gait dataset [38]. We carefully select 1,200 subjects all of them in the thirties age group. Each subject performs a complete walking sequence under the lab environment. We extract the binary silhouette sequences for all subjects. Afterward, we hired seven annotators to watch the binary sequences for each pair of subjects and record their observations as mentioned in Table 1 in term of attribute assignment (1,0, or -1).

For each gait video, we extract the dense trajectories, and for each DT, we assign 3D volume around it. The 3D volume has the spatial-temporal dimension $16 \times 16 \times 15$. We divide the volume into 12 sub-blocks, each one has the dimension $8 \times 8 \times 5$. For each sub-block, we compute the mean-speed subtracted HOF (9-bin orientation quantization) then normalized it by its **L1**-norm followed by a square root. All the sub-blocks HOFs are then concatenated into single DT-HOF. Now we have DT-HOF descriptor for each dense trajectory. As the number of extracted DTs-HOFs are large for each gait video, we adopt the Fisher encoding mechanism to encode the motion information of the entire video into a single global feature vector. Following the formulation in 3.3, we select the corresponding DTs-HOFs of 800 subjects from the dataset for training a GMM with 256 clusters. We then use the estimated parameters to compute the FVs for all 1,200 subjects as explained. For now, we have a single FV for each video. This FV only encodes the motion information for each subject (i.e., we have not yet consider the appearance information).

5.1 Zero-shot Learning Results

For training, we use the Rank-SVM classifier which is adapted by Parikh et al. [25] to handle the relative attributes instead of binary labels. We completely disjointed the dataset into 1000 FVs for training and 200 FVs for testing. We use the training set to build the ordered and un-ordered pairs sparse matrices which are needed for classifier. The output of classifier is the zero-shot weights. Then, we use these estimated weights to compute a set of ranking functions for training and testing (totally unseen)

Table 2: The attribute estimation accuracies for seven different annotators. Both the training and testing phases accuracies are reported. The average accuracy for each annotation is mentioned in the rightmost column.

Subjects Pairs		Annotators	General goodness	Stately	Cool	Relax	Arm swing	Step length	Walking speed	Spine	Average Accuracy(%)
1,200	Training	Annotator 1	66	78	95	75	73	81	73	75	77
		Annotator 2	71	72	73	75	77	76	72	72	73
		Annotator 3	83	70	68	75	83	80	74	82	77
		Annotator 4	75	76	71	74	76	80	68	75	74
		Annotator 5	72	65	68	62	75	79	75	75	71
		Annotator 6	77	78	68	69	79	74	77	74	75
		Annotator 7	70	68	74	80	67	66	75	70	71
	Testing	Annotator 1	56	73	85	68	69	82	67	77	72
		Annotator 2	55	49	64	70	62	81	72	48	63
		Annotator 3	65	57	42	71	81	74	72	64	66
		Annotator 4	54	51	52	72	51	61	60	54	57
		Annotator 5	49	44	61	52	63	64	67	48	56
		Annotator 6	73	71	49	62	64	69	66	58	64
		Annotator 7	62	57	57	74	57	54	59	52	59

samples using Eq. (4). We finally, use the resulted real-valued ranks for evaluating the classifier against the GT annotation using the evaluation criteria that proposed in 4.2. Note that, we assume multi-thresholds evaluation and examine the proposed method performance under the individual annotations. We use the seven gait attributes annotations individually and measure the classification accuracy of the predicted attributes for each annotator. The quantitative results are listed in Table 2. several observations we realize from the results. Firstly, as the pairwise gait attributes annotations depends on human perception, so different annotators may report different labels of the same attribute and this is, in turn, influences the estimation process. The proposed method adapts itself properly to the human perception of the annotators. Secondly, the experimental results uncover the interesting observation that the gait style has key features attributes that influence the annotator decision. In Table 2, we observe higher accuracy rates for all annotators at *arm swing*, *step length*, and *walking speed* attributes. It means that although each annotator performed the annotation separately without biasing, they almost agreed that these attributes have a strong presence in subjects gait styles. This observation may lead gait community to look deeply inside the persons' walking patterns and focus on the discriminative features that best describe the gait style. At a high semantic level of analysis, the social relationships between the subjects may be investigated based on estimated relative attributes. As well, The human perception itself of the annotators can be inferred. To measure the robustness of our DT-based representation for gait attribute estimation, we compare it against the GEI-based deep features representation under the same classifier and evaluation criteria setups. We use the VGG16 deep architecture [31] which produces a 4096-dimensional feature descriptor from the walking person's GEI. It worth noting that GEI encodes both the appearance and motion information of the walking subject in contrast to our representation which carries only the motion information yet. The evaluation results based on the GEI-based feature representation are listed in Table 3 in the first and third columns. We can observe that

Table 3: The accuracies of gait attributes estimation based on our proposed DT-based representation versus GEI-deep based representation.

Attributes)		General goodness	Stately	Cool	Relax	Arm swing	Step length	Walking speed	Spine	Average Accuracy(%)
Training	GEI features	**66.08**	**62.36**	**62.72**	**60.93**	**62.93**	**64.17**	**61.50**	**68.46**	**63.56**
	DTs features	63.62	60.28	60.37	58.32	61.89	63.27	60.41	65.20	61.67
Testing	GEI features	**55.89**	**48.08**	52.86	47.92	46.01	58.59	47.61	50	50.87
	DTs features	55.41	47.61	**57.96**	**56.84**	**58.75**	**61.30**	**61.78**	48.56	**56.03**

GEI-based representation slightly outperforms our representation at the training stage. By contrast, in the testing stage, our DT-based representation outperforms the GEI-based representation at most accuracy scores with remarkable enhancements. Sample results based on our proposed representation appear in Table 3 in the second and fourth rows (highlighted by yellow color) and the average accuracy for all attributes appears in the rightmost column. In this experiment, we concatenate all the seven annotations which produce 8389 subjects pairs after excluding the unrecognizable pairs.

6 Conclusion

In this paper, we exploit the concept of learning to rank in the human gait attributes estimation for the first time. Instead of using the traditional binary classification problem, we directly learn a set of ranking functions based on the preference relationships between walking subjects pairs. We consider each walking subject has a gait style represented by a set of relative attributes. We propose a novel pairwise gait attributes annotations for 1,200 walking subjects based on human perception. In this work, we propose to encode only the gait motion information of subjects based on dense trajectory-based representation. We build global gait motion descriptors for the walking subjects based on the HOFs descriptors of the extracted DTs. For gait attribute prediction, we learn a set of ranking functions from training samples given the proposed annotations. The initial results show the robustness of the proposed DTs-based gait motion representation compared to the GEI-deep features representation which encodes both the appearance and motion information.The experimental results showed that the proposed method could well represent gait attributes and also that the proposed gait motion descriptor had better generalization capability than GEI for gait attributes estimation task.

Acknowledgement

This work was supported by JSPS Grants-in-Aid for Scientific Research (A) JP18H04115.

References

1. Akae, N., Mansur, A., Makihara, Y., Yagi, Y.: Video from nearly still: An application to low frame-rate gait recognition. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 1537–1543. IEEE (2012)

2. Bobick, A., Johnson, A.: Gait recognition using static activity-specific parameters. In: Proc. of the 14th IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 423–430 (2001)
3. Bouchrika, I., Nixon, M.S.: Model-based feature extraction for gait analysis and recognition. In: International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications. pp. 150–160. Springer (2007)
4. Castro, F.M., Marín-Jiménez, M.J., Mata, N.G., Muñoz-Salinas, R.: Fisher motion descriptor for multiview gait recognition. International Journal of Pattern Recognition and Artificial Intelligence **31**(01), 1756002 (2017)
5. Castro, F.M., Marín-Jimenez, M.J., Medina-Carnicer, R.: Pyramidal fisher motion for multiview gait recognition. In: Pattern Recognition (ICPR), 2014 22nd International Conference on. pp. 1692–1697. IEEE (2014)
6. Crammer, K., Singer, Y.: Pranking with ranking. In: Advances in neural information processing systems. pp. 641–647 (2002)
7. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1785. IEEE (2009)
8. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. pp. 363–370. Springer (2003)
9. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE transactions on pattern analysis and machine intelligence **28**(2), 316–322 (2006)
10. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Advances in neural information processing systems. pp. 487–493 (1999)
11. Kusakunniran, W.: Attribute-based learning for gait recognition using spatio-temporal interest points. Image and Vision Computing **32**(12), 1117–1126 (2014)
12. Kusakunniran, W.: Recognizing gaits on spatio-temporal feature domain. IEEE Transactions on Information Forensics and Security **9**(9), 1416–1423 (2014)
13. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 951–958. IEEE (2009)
14. Laptev, I.: On space-time interest points. International journal of computer vision **64**(2-3), 107–123 (2005)
15. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies (2008)
16. Liu, T.Y., et al.: Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval **3**(3), 225–331 (2009)
17. López-Fernández, D., Madrid-Cuevas, F.J., Carmona-Poyato, Á., Marín-Jiménez, M.J., Munoz-Salinas, R.: The ava multi-view dataset for gait recognition. In: International Workshop on Activity Monitoring by Multiple Distributed Sensing. pp. 26–39. Springer (2014)
18. Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y.: Gait recognition using a view transformation model in the frequency domain. In: Proc. of the 9th European Conference on Computer Vision. pp. 151–163. Graz, Austria (May 2006)
19. Makihara, Y., Mannami, H., Yagi, Y.: Gait analysis of gender and age using a large-scale multi-view gait database. In: Asian Conference on Computer Vision. pp. 440–451. Springer (2010)
20. Makihara, Y., Matovski, D.S., Nixon, M.S., Carter, J.N., Yagi, Y.: Gait Recognition: Databases, Representations, and Applications, pp. 1–15. John Wiley & Sons, Inc. (1999). https://doi.org/10.1002/047134608X.W8261, `\url{http://dx.doi.org/10.1002/047134608X.W8261}`

21. Makihara, Y., Okumura, M., Iwama, H., Yagi, Y.: Gait-based age estimation using a whole-generation gait database. In: 2011 International Joint Conference on Biometrics (IJCB). pp. 1–6. IEEE (2011)
22. Marín-Jiménez, M.J., de la Blanca, N.P., Mendoza, M.A.: Human action recognition from simple feature pooling. Pattern Analysis and Applications **17**(1), 17–36 (2014)
23. Marín-Jiménez, M.J., Castro, F.M., Carmona-Poyato, Á., Guil, N.: On how to improve tracklet-based gait recognition systems. Pattern Recognition Letters **68**, 103–110 (2015)
24. Nordin, M., Saadoon, A.: A survey of gait recognition based on skeleton mode l for human identification. Research Journal of Applied Sciences, Engineering and Technology (2016)
25. Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision. pp. 503–510. IEEE (2011)
26. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
27. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. Computer Vision–ECCV 2010 pp. 143–156 (2010)
28. Rida, I., Almaadeed, N., Almaadeed, S.: Robust gait recognition: a comprehensive survey. IET Biometrics **8**(1), 14–28 (2019). https://doi.org/10.1049/iet-bmt.2018.5063
29. Rida, I., Al Maadeed, N., Marcialis, G.L., Bouridane, A., Herault, R., Gasso, G.: Improved model-free gait recognition based on human body part. In: Biometric Security and Privacy, pp. 141–161. Springer (2017)
30. Shashua, A., Levin, A.: Ranking with large margin principle: Two approaches. In: Advances in neural information processing systems. pp. 961–968 (2003)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Sugimura, D., Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Using individuality to track individuals: clustering individual trajectories in crowds using local appearance and frequency trait. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 1467–1474. IEEE (2009)
33. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision **2**(5), 2–2 (2002)
34. Uijlings, J., Duta, I.C., Sangineto, E., Sebe, N.: Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. International Journal of Multimedia Information Retrieval **4**(1), 33–44 (2015)
35. Uijlings, J.R., Duta, I.C., Rostamzadeh, N., Sebe, N.: Realtime video classification using dense hof/hog. In: Proceedings of international conference on multimedia retrieval. p. 145. ACM (2014)
36. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action recognition by dense trajectories (2011)
37. Weng, J., Lu, W., Xu, J., et al.: Multi-gait recognition based on attribute discovery. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
38. Xu, C., Makihara, Y., Ogi, G., Li, X., Yagi, Y., Lu, J.: The ou-isir gait database comprising the large population dataset with age and performance evaluation of age estimation. IPSJ Transactions on Computer Vision and Applications **9**(1), 24 (2017)
39. Yan, C., Zhang, B., Coenen, F.: Multi-attributes gait identification by convolutional neural networks. In: 2015 8th International Congress on Image and Signal Processing (CISP). pp. 642–647. IEEE (2015)
40. Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S.: Robust relative attributes for human action recognition. Pattern Analysis and Applications **18**(1), 157–171 (2015)