Deep Gait Relative Attribute using a Signed Quadratic Contrastive Loss

Yuta Hayashi*, Allam Shehata*†, Yasushi Makihara*, Diago Muramatsu‡* and Yasushi Yagi*
* The Institute of Scientific and Industrial Research, Osaka University, Japan
Email: {hayashi, allam, makihara, muramatsu, yagi}@am.sanken.osaka-u.ac.jp
† Informatics Department, Electronics Research Institute, Egypt
Email:allam@eri.sci.eg
‡ Faculty of Science and Technology, Seikei University, Japan
Email:muramatsu@st.seikei.ac.jp

***Abstract*—This paper presents a deep learning-based method to estimate gait attributes (e.g., stately, cool, relax, etc.). Similarly to the existing studies on relative attribute, human perception-based annotations on the gait attributes are given to pairs of gait videos (i.e., the first one is better, tie, and the second one is better), and the relative annotations are utilized to train a ranking model of the gait attribute. More specifically, we design a Siamese (i.e., two-stream) network which takes a pair of gait inputs and output gait attribute score for each. We then introduce a suitable loss function called a signed contrastive loss to train the network parameters with the relative annotation. Unlike the existing loss functions for learning to rank does not inherit a nice property of a quadratic contrastive loss, the proposed signed quadratic contrastive loss function inherits the nice property. The quantitative evaluation results reveal that the proposed method shows better or comparable accuracies of relative attribute prediction against the baseline methods.**

I. Introduction

Walking is the most fundamental action for human beings and also a primary mode of locomotion. We can observe a variety of information from his/her walking style, i.e., gait, which include but not limited to identity [2], [4], [6], [8], [11], [17], [36], [43], [61], [70], gender [32], [38], [71], age [25], [31], [34], ethnicity [72], disease [23], [27].

Besides, psychological researchers indicated that human gait has several attributes in [42], and the other subsequent studies also find several gait attributes [29], [39] ranging from perceptual ones (e.g., relaxed vs. nervous, happy vs. sad, etc.) to physical ones (arm length, arm swings, etc.). For example, let's look at gait silhouette sequences of two subjects in Fig. 1. We may perceive differences in several gait attributes; subject A walks more actively and stately, which may be caused by wider stride, larger arm swing, less stoop, etc.

If a machine as well as a human can perceive such kinds of gait attributes, it can contributes to several applications of attractiveness computing, i.e., a person may be able to know the degree of attractiveness of his/her gait in an objective manner and improve it so as to increase the gait attractiveness, or a gait instructor may use it similarly. In addition, the other perceptual and physical gaits attributes other than the above-mentioned attractiveness, have also a variety of potential applications such as criminal investigation with a witness (e.g., a witness may provide a couple of attributes: the degree of arm swing, the degree of stoop, etc.), and a health care application (e.g., depression may be early detected by continuously observing emotional gait attribute). In fact, instead of the existing holistic appearance-based gait features, some automatically extracted gait attributes (e.g., walking speed, step length, arm swing, etc.) [7] have been already employed to enhance gait recognition accuracy in [7], [20], [68].

Figure 1. Gait silhouette sequences from two subjects (white: person, gray: cast shadow, black: background). We may find that subject A has relatively wider stride, larger arm swing, less stoop than subject B, and his gait may be perceived to be generally good, cool, etc.

In order to make the machine perceive the gait attribute, we naturally need to collect annotation data (i.e., gait video and its corresponding attribute values) similarly to general machine learning problems. We, however, face difficulty in annotating the gait attributes, unlike we can clearly annotate subject IDs for gait recognition, and gender/ages for gait-based gender/age estimation. For example, let's assume that you are asked to annotate a subject A's gait attribute (e.g., general goodness) with the gait silhouette sequence shown. You may perceive that it is somewhat generally good, but may feel unconfident in giving it an absolute score (e.g., 85 points out of 100 points).

Because this is often the case with studies on attributes not only on gait but also face, object, etc., an alternative annotation and machine learning schemes has been proposed by Parikh and Grauman [45], which is so-called *relative attribute*. In the relative attribute framework, instead of annotating absolute scores for each of training samples, a pair of training samples are shown to an annotator and he/she gives a relative score (e.g., the first one is better, similar, or the second one is better). For example in Fig. 1, we conjecture that people are confident in annotating "subject A is better" in terms of the general goodness of gait. Techniques of machine learning to ranking

(e.g., ranking support vector machine [16]) are then applied to the annotated training samples and then obtain an attribute estimator for each sample, although the estimated attribute is still relative one and hence up to scale and translation (bias).

The relative attribute frameworks have been recently extended to deep learning ones [54], [69] similarly to other computer vision and pattern recognition tasks, where feature extraction and learning to rank are simultaneously optimized in an end-to-end way. While the deep relative attributes have been already introduce in face and object attributes [54], [69], they have not yet in the gait attributes. The existing studies on the relative gait attribute [1], [39], [51] are still based on traditional machine learning to ranking, and hence it is expected to improve the accuracies of estimating the gait relative attributes by introducing the deep learning framework.

We therefore introduce a deep relative gait attribute framework for the first time in this paper. More specifically, we aim at the gait attributes on human perception-based attractiveness (e.g., general goodness, cool, stately, relaxed, etc.) similarly to [51] given a silhouette-based gait template as an input.

Main contributions of this work are summarized as follows.

1) **The first deep learning framework to estimate relative gait attribute**
   We introduce a deep relative gait attribute framework for the first time. Unlike the existing methods [38], [51] extract features and learning to ranking as different steps, the proposed method realize it in an end-to-end manner thanks to the deep learning framework.
2) **A signed quadratic contrastive loss function for effective learning to rank**
   In order to learn relative attribute (the first one is better, similar, the second one is better), it is essential to consider a sign of attribute score difference of an input pair (i.e., the first one's score minus the second one's score). While existing methods on deep relative attributes [54], [69] employ loss functions whose gradient does not change in proportion to the score difference, we propose a signed quadratic contrastive loss function whose gradient changes in proportion to the score difference, which makes learning process more effective.

II. Related work

A. Gait recognition

Recently, recognizing people by their gait (i.e., walking style) has become most popular. This is because gait biometric is attractive can be captured remotely without special attached instruments or subject's participation [58]. Moreover, the gait biometric is typically on-invasive biometric, which make it hard to spoof and reliable for various gait recognition, identification, and verification applications [33], [35], [48], [49]. Currently, there are several publicly available large population gait datasets [30], [67], [76]. As a result, the availability of these datasets opened the doors to tackle many gait recognition challenging tasks and proposing several gait analysis/recognition algorithms with promising performance. [35], [37], [55], [65]. There are massive number of gait recognition methods that have been proposed and published in last two decades. We refer the reader to the recent surveys [14], [58]. There are two main types of gait recognition approaches (1) video-based, where a camera is used to capture the gait sequence and (2) sensor-based approaches, where sensors (e.g. typically, accelerometers and gyroscopes) are attached to the human body [58]. Because of its flexibility, vision-based approach is widely used where only camera(s) used to capture the gait data. Furthermore,from the modeling viewpoint, two types of video-based approaches have been founded, namely model-free and model-based approaches. The model-based approaches typically build a predefined model of the kinematics of human joints to measure a set of physical gait parameters such as step length and angular speed [3], [28], [60]. On the other hand, model-free methods directly extracts features from silhouette frames and the human gait is represented compactly as a whole without knowing the underlying structure of the human body [44], [56], [58]. For instance, Han and Bhanu have proposed the GEI descriptor [11] to encode the gait features. It is the most prevalent and frequently used feature in the recent gait recognition systems [6], [7], [52].

B. Attribute

Generally speaking, traditional action recognition systems, including gait, directly associate low-level features with class labels. However, discriminative visual descriptions cannot be characterized easily using only a single class label. Therefore, to describe action-related properties, high-level semantic concepts should be considered [75]. Recently, attribute-based methods have demonstrated the usefulness of object attributes for high-level semantic description, functioning well for action recognition (including gait), video classification, and zero-shot learning [10], [21], [22], [45].

C. Relative attribute

Currently, the visual attributes terminology is adopted to principally refer to those perceptual properties that can be used in a more spontaneous manner to describe the visual entities (i.e. image, scene, or object). These perceptual properties often shared among visual entities categories. Because of it is reliable, human understandable, and easy to detect automatically [50], visual attributes have been widely used in many applications such as object detection/description [57], [63], [64], image retrieval [18], [74] and transfer learning [24], [26], [46]. Most of the existing attributes-based learning approaches consider the attributes as hard-wired representation (i.e., binary), to describe the existence (or absence) of the perceptual property. This is may make it hard for such approaches to deal with describing/detecting the unseen entities. Moreover, several classification-based systems directly associate low-level features with absolute annotation labels. Even though, it might be more appropriate to define high-level semantic concepts in terms of relative attribute labels instead of absolute labels. Recent research indicates that these attributes can be adopted in human recognition tasks as an intermediate level

of description for human visual properties [75]. To handle this issue, Parikh et al. [45] have defined the *relative attribute* notion and used it to learn a set of relative attributes-based ranking functions from images/objects. Those learned ranking functions are then used to relate the unseen images/objects to the training images and predict the relative strength of each attribute. Using relative attributes, image/object can be described semantically in such a way that not only reports the existence of an attribute but also predict the strength of that attribute.

Following the concept of relative attributes and inspired by the success of convolutional neural networks, Tang et. al. [69] and Souri et. al. [54] introduced the two recent deep relative attributes (DRA) frameworks to learn the visual properties from the single input image and use effective nonlinear ranking functions to describe the relative attributes among the image pairs. The authors also formulated different relative loss functions to constrain the predicted relative attributes' strengths for the ordered pairs (i.e., one is better than the other) and the unordered image pairs (i.e., tie). Furthermore, relative attributes started to be utilized widely and combined with the deep learning frameworks. The authors in [73] proposed to utilize the acquired attribute-correlated local regions of the image. They do not adopt discovering the spatial extent of the corresponding attribute based on the ranking list of all the images in the dataset. Instead, they classify the images according to the existence of the attribute. A robust approach is proposed in [66] to discovers the spatial extent of relative attributes. A novel formulation is introduced to combines a detector with local smoothness to discover chains of visually coherent patches, efficiently generate additional candidate chains, and rank each chain according to its relevance to the attribute. The authors in [9] propose to benefit from the deep learning models' capabilities to characterize the nature of the relationships learned by the CNN model and establish mid-level representations used for semantic visual attributes. A visual Attribute-augmented multi-stream 3D CNN framework proposed in [64] for action recognition. The visual attribute pipeline includes an object detection network, an attributes encoding network and a classification network. To handle the problems of directly exploiting pairwise comparisons and the ignorance of the relationship among different attributes, a multi-task deep relative attribute learning model based on Siamese networks is proposed in [41]. This model is capable of learning all the perceptual attributes simultaneously via multi-task learning. Moreover, a weakly supervised method is proposed in [62] for simultaneously learning outdoor scene parts and attributes from set of images associated with attributes in the text while the precise localization of each attribute left unknown.

D. Attributes in gait analysis

The visual attributes concept has been recently utilized in gait analysis to improve the performance. In [20] the attribute-based classification is applied for gait recognition enhancement by reducing the classifier models needed for recognizing each probe gait. This process significantly reduces the computational complexity at the testing phase as well as improve the recognition accuracy. The authors in [68] used a deep learning model combined with a multi-task learning model to identify human gait and predict the gait attributes simultaneously. A novel method of human description is proposed in [47] based on set of human soft attributes. On the other hand, an attribute discovery model is proposed in [7] for multi-gait recognition. Using the extracted GEI, stable and discriminative attributes are developed using a latent conditional random field (L-CRF) model. In the recognition process, the attribute set of each person is detected by inferring on the trained L-CRF model. For person identification, a set of semantic clothing attributes jointly with human body parts is considered in [15]. Inspired by the relative attributes, a super-fine attributes concept is introduced in [40] to discover more relevant and precise human description used for person re-identification. Although, it became easier to train models and use them to identify a person based on his/her gait, it still difficult for such models to recognize the gait of persons who have never been seen before, or at least, relate them to the observed ones based on their gait relative attributes [51]. Therefore, we introduce a deep relative gait attribute framework for the first time. Unlike the existing methods [38], [51] which adopt feature extraction and learning to ranking as different steps, the proposed method realize it in an end-to-end manner.

III. Deep relative gait attribute

A. Problem setting

In our gait attribute estimation framework, given a gait template $\boldsymbol{x}$ (e.g., GEI [11]), we estimate its corresponding score r of a certain gait attribute (e.g., the degree of general goodness) as

$$r = f(\boldsymbol{x}), \tag{1}$$

where $f(\cdot)$ is a mapping function from a gait template to an attribute score.

Because it is difficult to directly annotate the attribute score r for training, we rely on the relative attribute framework in this work as discussed in the introduction section. Specifically, we show a pair of videos of two different walking persons to an annotator and then ask him/her to give a comparative label y to the pair with regard to a certain gait attribute (e.g., general goodness of the gait). The comparative label y is chosen from a ternary set $\mathcal{Y} = \{1, 0, -1\}$: $y = 1$ when the first one is better; $y = 0$ when the first and second ones are similar; and $y = -1$ when the second one is better. Note that a pair with label $y = 0$ is so-called an unordered pair, while a pair with label $y = 1$ or $y = -1$ is ordered pair.

Once we collect N pairs of the gait templates and its corresponding comparative label as $\{\boldsymbol{x}_{1,i}, \boldsymbol{x}_{2,i}, y_i\}(i = 1, \ldots, N)$ as a result of annotation, we training the mapping function f so as that a score difference of the i-th pair

$$d_i = r_{1,i} - r_{2,i} = f(\boldsymbol{x}_{1,i}) - f(\boldsymbol{x}_{2,i}) \tag{2}$$

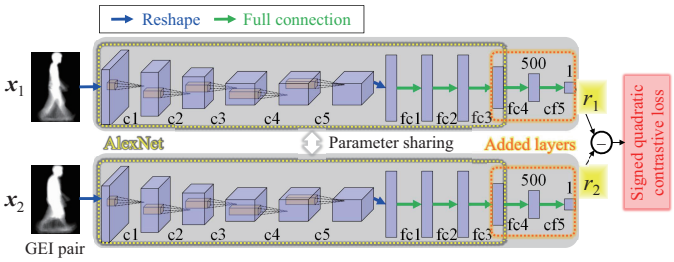


Figure 2. Network architecture.

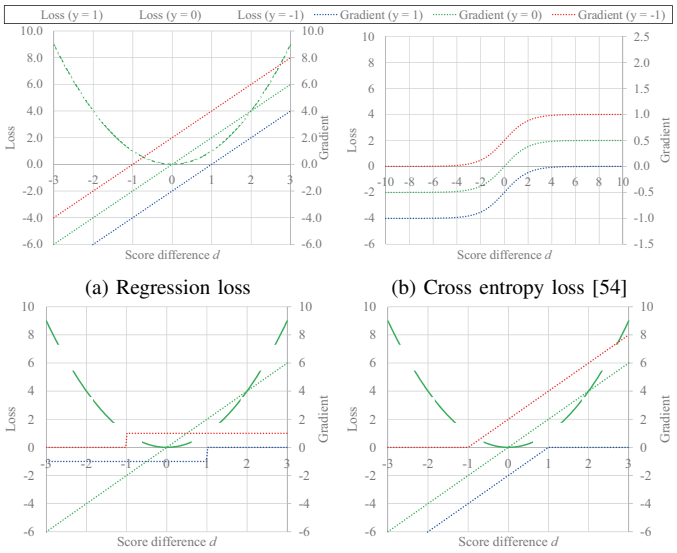
can be consistent with the i-th comparative label y_i as much as possible. Intuitively speaking, if the comparative label $y_i = 1$ (i.e., the first one is better), it is preferable that the score of the first one $r_{1,i}$ is larger than that of the second one $r_{2,i}$ and hence the score difference d_i should be positive. If the comparative label $y_i = -1$ (i.e., the second one is better), the score difference d_i should be negative similarly. If the comparative label $y_i = 0$ (e.g., the first and second ones are similar), it is ideal that the score of the first one is the same as that of the second one, and hence the score difference d_i should be close to 0.

We describe how to train the mapping function f using a deep learning framework with a loss function particularly designed for learning to rank.

B. Network architecture

Similarly to the existing approaches to deep relative attribute [54], [69], we employ a Siamese network composed of two streams, which takes a pair of gait templates as an input and output a corresponding pair of attribute scores as shown in Fig. 2. Since the parameters are shared between the two streams, we can feed a single gait template to get its attribute score in a test stage, i.e., we need not to feed a pair of gait templates, although the obtained attribute score is up to scale and translation (bias).

Specifically, we employ a GEI as the input gait template as it has been used for a long time in video-based gait analysis research community due to its simple yet effective representation. In addition, the network is composed of conventional layers such as convolution, normalization, pooling, and full connection layers and is configured based on AlexNet [19], which is one of standard network structures in computer vision and pattern recognition community. We made a slight modification from the original AlexNet in the last layers, i.e., we added two extra fully connected layers inspired by [69] where the last output of ours is a scalar value (or a single node) which indicates an attribute score, while the output of the original is a multi-dimensional nodes of object classes. Note that it is not a scope of this paper to design a technically novel network architecture. In other words, we can flexibly replace it with other backbone networks such as VGG [53], ResNet [12], and DenseNet [13].



(a) Regression loss (b) Cross entropy loss [54]

(c) Signed linear contrastive loss [69] (d) Signed quadratic contrastive loss

Figure 3. Loss and its gradients for ranking loss functions. Note that the loss is scaled with the primary vertical axis (at the left side), while the gradient is scaled with the secondary vertical axis (at the right side). A margin m is set to 1 in (c)(d).

C. Loss functions

A loss function is crucial for learning to ranking, and hence we focus on the loss function aspect in this paper. As we discussed at the beginning of this section, we cannot directly regress the attribute score for each sample and hence consider the consistency between the score difference $d_i = r_{1,i} - r_{2,i}$ and the comparative label $y_i \in \mathcal{Y}$. We introduce our proposed loss function by addressing drawbacks of the existing loss functions one by one.

1) Regression loss (see Fig. 3(a)): The most straightforward way is to directly regress the score difference to the comparative label with the following regression loss L_{reg}.

$$L_{\mathrm{reg}} = \sum_{i=1}^{N} (d_i - y_i)^2. \tag{3}$$

While the regression loss is reasonable for unordered pairs (i.e., $y_i = 0$), it does not ideally behave for ordered pairs (i.e., $y_i = 1$ or $y_i = -1$) as also addressed in [54]. Here, let's think of the case $y_i = 1$, where the first one is better ($d_i = r_{i,1} - r_{i,2} > 0$). Taking it into consideration the meaning that the first one is better, if the score difference d_i get larger than a certain positive value (e.g., $d_i > 1$), it should not be penalized. For example, if the loss value is zero for the score difference $d_i = 1$, the loss value should be less than or equal to zero for the larger score difference (e.g., $d_i = 2$). The regression loss, however, penalizes the score difference when it gets larger than 1 by definition with Eq. (3) also as shown in Fig. 3(a).

D. Cross-entropy loss (see Fig. 3(b))

Souri et al. [54] proposed another loss function for learning to rank to overcome the drawback of the above-mentioned regression loss. More specifically, they cast this problem as a two-class classification problem, where the first class means that the first one is better ($y_i = 1$), and the second class means that the second one is better ($y_i = -1$). In the framework, they consider two probabilities: the probability $p_{1,i}$ for the first class; and the probability $p_{2,i}(= 1 - p_{1,i})$ for the second class, and then represent a similar case ($y_i = 0$) by an equilibrium status, i.e., ($p_{i,1} = p_{i,2} = 0.5$). They therefore introduce an auxiliary label t_i to represent $p_{i,1}$ and its actual correspondence is as follows: $t_i = 1$ for $y_i = 1$; $t_i = 0.5$ for $y_i = 0$; and $t_i = 0$ for $y_i = -1$. They convert a score difference d_i into the probability $p_{1,i}$ by a logistic function as

$$p_{1,i}(d_i) = \frac{1}{1 + e^{-d_i}}. \tag{4}$$

They finally compute a cross-entropy loss between the estimated probability distribution and the auxiliary label as

$$L_{\mathrm{CE}} = -\sum_{i=1}^{N} \{t_i \log p_{1,i}(d_i) + (1 - t_i) \log(1 - p_{1,i}(d_i)\}. \tag{5}$$

As depicted in Fig. 3(b), we can see that the larger score difference is not penalized unlike the regression loss (Fig. 3(a)).

On the other hand, taking a closer look at the gradient, we notice that the gradient hardly changes in certain domains. Here, let's think of the case $y_i = 1$ again, where the first one is better ($d_i = r_{i,1} - r_{i,2} > 0$). We further assume that there are two pairs whose estimated score differences are inconsistent with the comparative label: one is -5 and the other is -10. Since the degree of violation for the latter pair is as double as the former pair, we should weigh the latter pair more for faster convergence. There is, however, almost no gradient difference between them (gradients are approximately -1.0 for the both sample), and hence the cross-entropy loss with logistic approximation may not be necessarily the optimal choice.

E. Signed linear contrastive loss (see Fig. 3(c))

Yang et al. [69] introduced another loss function which is inspired by a contrastive loss typically used for two-class classification problem given a pair, i.e., judge whether the pair belongs to the same class ($\delta_i = 0$) or different classes ($\delta_i = 1$). The original contrastive loss takes a non-negative dissimilarity (or distance) D_i as an input, and tries to make it as small as possible for the same class label, while keep it away each other than a certain margin m. The loss for the ith sample can be written as

$$L_{\mathrm{C},i} = \begin{cases} D_i^2 & (\delta_i = 0) \\ \max(0, m - D_i)^2 & (\delta_i = 1) \end{cases}. \tag{6}$$

The contrastive loss in total is then written as

$$L_{\mathrm{C}} = \sum_{i=1}^{N} \{(1 - \delta_i) D_i^2 + \delta_i \max(0, m - D_i)^2\}. \tag{7}$$

For learning to rank, we need to handle not a non-negative dissimilarity D_i but a signed difference d_i, they modify the above-mentioned contrastive loss for the i-th sample to handle the sign as

$$L_{\mathrm{SLC},i} = \begin{cases} d_i^2 & (y_i = 0) \\ \max(0, m - d_i) & (y_i = 1) \\ \max(0, m + d_i) & (y_i = -1) \end{cases}. \tag{8}$$

Note that the score difference $d_i = r_{1,i} - r_{2,i}$ should be larger than the margin m in case of $y_i = 1$ (i.e., the first one is better), while that it should be smaller than the minus margin $-m$ in case of $y_i = -1$ (i.e., the second one is better). We may refer the readers to Fig. 3(c) for better intuitive understanding of the behavior of this loss function. The signed linear contrastive loss in total is then summarized as

$$L_{\mathrm{SLC}} = \sum_{i=1}^{N} \left\{(1 - |y_i|) d_i^2 + |y_i| \max(0, m - y_i d_i)\right\}. \tag{9}$$

We notice that the loss function is a quadratic form for an unordered pair ($y_i = 0$), while it is a linear form for an ordered pair ($y_i = 1$ or $y_i = -1$), which results in significant difference in gradient magnitude and also in convergence property between the unordered pairs and the ordered pairs. Besides, because the gradient for the ordered pairs does not change regardless of the degree of violation, this loss function suffers from the same problem as discussed in the previous subsection.

1) Signed quadratic contrastive loss (see Fig. 3)(d): In order to solve the above-mentioned inconsistent treatment between the unordered pairs and the ordered pairs as well as the gradient unchanging problem in certain domains, we introduce a quadratic version of the signed contrastive loss function. In order to make the loss function a quadratic form by keeping the sign for the ordered pairs, we multiply the absolute score difference for each of the cases $y_i = 1$ and $y_i = -1$, and we formulate the loss function for the i-th sample as

$$L_{\mathrm{SQC},i} = \begin{cases} d_i^2 & (y_i = 0) \\ \max\{0, (m - d_i)|m - d_i|\} & (y_i = 1) \\ \max\{0, (m + d_i)|m + d_i|\} & (y_i = -1) \end{cases}. \tag{10}$$

The signed quadratic contrastive loss in total is then summarized as

$$\begin{aligned} L_{\mathrm{SQC}} \quad = \quad & \sum_{i=1}^{N} \big[(1 - |y_i|) d_i^2 \\ & + |y_i| \max\{0, (m - y_i d_i)|m - y_i d_i|\}\big]. \end{aligned} \tag{11}$$

Thanks to the signed quadratic property, we can weigh a sample whose degree of violation is large because the magnitude of the gradient increases in proportion to the score difference as shown in Fig. 3(d) unlike the cross-entropy loss and the signed linear contrastive loss functions, and hence we expect the better convergence property with the signed quadratic contrastive loss.

IV. Experimental evaluation

A. Dataset

To the best of our knowledge, the first gait relative attributes dataset has been released in [51]. This dataset is compiled from the publicly available gait recognition dataset, OULP-Age [67]. A set of 1,200 subjects' walking videos were selected from this dataset. These walking videos were then arranged into pairs of subjects and presented to several annotators (i.e. six) for comparative annotation. In total, 1,200 pairs have been selectively generated.

Eight gait attributes have been defined as $\{$*General goodness, Stately, Cool, Relaxed, Arm swing, Walking speed, Step length, Spine*$\}$. Each of these attributes describes a certain visual property of the walking subject and could receive the comparative labels from $\mathcal{Y}$. The annotators were instructed to assign $y = 1$ to the gait attribute if they observed that the attribute's strength in the gait style of the first subject is greater than that of the second subject, assign $y = -1$ vice versa. As well, if the annotators judged that both subjects have the same attribute's strength, they report the label $y = 0$.

B. Experimental setup

From the total of 1,200 pairs, we extracted 900 training pairs and 100 test pairs. Note that each subject appears twice in 1,200 pairs in this dataset and hence the remaining 200 pairs were discarded to keep the subjects in the training and test sets completely disjoint. The dataset contains the comparative labels for each pair from six annotators. Therefore, we adopt the majority voting when the comparative labels are inconsistent. As well, we excluded the tie score pairs[1] from both the training and testing pairs so that the training and testing pairs used in our experiments differ from one gait relative attributes to another.

We trained all the network parameters using stochastic gradient decent algorithm [5] with a mini batch size of 64. The learning rate was set to 10^{-3} during 200 epochs and the average loss over the batch is computed. The margin m for the signed linear/quadratic contrastive loss is set to 1. The network parameters are trained for each gait attribute separately.

Since the annotation for the test set is also given in the form of the comparative labels, accuracy evaluation is also done in a pairwise way. For each test pair, we compute a score difference d and then classify into three classes. Specifically, we set a predictive comparative label $\hat{y}$ as follows: the first one is better ($\hat{y} = 1$) if $d > d_{\mathrm{thresh}}$; the second one is better ($\hat{y} = -1$) if $d < -d_{\mathrm{thresh}}$; and the first and second ones are similar ($\hat{y} = 0$) otherwise, where d_{thresh} is a threshold for this three-class classification and is determined so as to maximize the classification accuracy.

C. Comparison with state-of-the-arts

To evaluate the effectiveness of the proposed model, we compare the closely related state of the art approaches [45], [51], [54], [69]. For [51] and [45] baselines, the R-SVM classifier is adopted to learns a linear ranking function for each gait relative attribute independently. The gait motion information is only preserved in [51] by utilizing the dense trajectories [59] and used for ranking function learning. One the other hand, the extracted deep visual features from GEI using the VGG16 pre-trained model [53] are used in [45]. For [54], and [69] baselines, two deep relative attributes models are proposed respectively to learn visual features and ranking functions. As well, cross-entropy and singed linear contrastive loss functions are introduced respectively for model training. We examined the performance of both baselines based on the state of the art GRA dataset [51]. For accuracy evaluation, the 3-class classification criteria described in Section IV-B is adopted. From the reported quantitative results in Table I, it is clear that the proposed model jointly with the new loss function show better (best) or comparable (second best) accuracy of gait relative attribute prediction against the baseline methods.

D. Ablation study on loss functions

To emerge the merits of the proposed signed quadratic contrastive loss function over the existing loss functions, we conduct an ablative study experiment. This experiment assumes using the same network architecture, AlexNet [19], and fixed experiment setup as reported above. For each loss function, we train the model for each gait relative attribute independently. For testing, the accuracy evaluation is also done in a pairwise way. For each test pair, we compute the score difference and then classify it into three classes. The quantitative results in terms of prediction accuracy of our proposed loss compared to the existing loss functions are reported in Table II accordingly. The most striking result to emerge from this experiment is that the proposed loss function quantitatively outperforms the existing loss functions under the same network architecture and experiment setup.

V. Conclusion

In this paper, we considered the relative attribute for gait recognition task. We proposed a gait relative attribute estimation model based on human perception. Inspiring by convolutional neural networks, the proposed model can learn attribute-specific visual features from the input GEIs and map these features into ranking scores about the relative strength values of the gait attribute in GEI pairs. For model training, we proposed a new loss function to properly handle both the contrastive and similar constraints of the GEI pairs and achieve the consistency. Our quantitative results reveal the comparable performance of the proposed model against the baselines methods in terms of relative attribute prediction accuracy on the recent GRA dataset. In the future work, we aim at enhancing the performance and compiling more annotations for GRA dataset.

Acknowledgments

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

[1]For a particular pair, 3 annotators voted with comparative label and the other 3 annotators voted with another comparative label.

Table I
COMPARISON WITH BENCHMARKS IN CLASSIFICATION ACCURACY [%]. BOLD AND ITALIC BOLD INDICATE THE BEST AND THE SECOND BEST ACCURACIES, RESPECTIVELY. THIS CONVENTION IS CONSISTENT THROUGHOUT THIS PAPER.

Method \ Attribute	General goodness	Stately	Cool	Relax	Arm swing	Step length	Walking speed	Spine	Average
DTs+FV+R-SVM [51]	34	43	59	***79***	71	74	**83**	46	61
GEI+VGG+R-SVM [45]	68	67	58	71	66	62	57	63	64
RankNet [54]	**86**	69	**72**	69	***72***	77	71	71	73
DRA [69]	81	***71***	67	75	69	**82**	72	**80**	***75***
Proposed	***81***	**71**	**69**	***79***	**72**	***81***	***72***	***79***	**76**

Table II
CLASSIFICATION ACCURACY [%] OF THE ABLATION STUDY OF LOSS FUNCTIONS UNDER THE SAME BACKBONE NETWORK OF ALEXNET [19]

Loss function \ Attribute	General goodness	Stately	Cool	Relax	Arm swing	Step length	Walking speed	Spine	Average
Regression loss	77	73	67	76	69	81	72	75	74
Cross-entropy loss	76	70	66	***77***	***72***	***82***	70	**82**	74
Signed linear contrastive loss	***81***	***71***	***67***	75	69	**82**	***72***	***80***	***75***
Signed quadratic contrastive loss	**81**	**71**	**69**	**79**	**72**	81	**72**	79	**76**

REFERENCES

[1] K. S. Abubacker and L. Indumathi. Attribute associated image retrieval and similarity reranking. In *2010 International Conference on Communication and Computational Intelligence (INCOCCI)*, pages 235–240. IEEE, 2010.

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

[3] C. BenAbdelkader, R. Cutler, and L. Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proceedings of Fifth IEEE international conference on automatic face gesture recognition*, pages 372–377. IEEE, 2002.

[4] A. Bobick and A. Johnson. Gait recognition using static activity-specific parameters. In *Proc. of the 14th IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 423–430, 2001.

[5] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

[6] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8126–8133, 2019.

[7] X. Chen, J. Xu, and J. Weng. Multi-gait recognition using hypergraph partition. *Machine Vision and Applications*, 28(1-2):117–127, 2017.

[8] D. Cunado, M. S. Nixon, and J. N. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003.

[9] V. Escorcia, J. Carlos Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1256–1264, 2015.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.

[11] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[13] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, July 2017.

[14] E. R. Isaac, S. Elias, S. Rajagopalan, and K. Easwarakumar. Trait of gait: A survey on gait biometrics. *arXiv preprint arXiv:1903.10744*, 2019.

[15] E. S. Jaha and M. S. Nixon. Soft biometrics for subject identification using clothing attributes. In *IEEE international joint conference on biometrics*, pages 1–6. IEEE, 2014.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

[17] A. Kale, A. Sundaresan, A. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa. Identification of humans using gait. *IEEE Transactions on image processing*, 13(9):1163–1173, 2004.

[18] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210, 2015.

[19] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[20] W. Kusakunniran. Attribute-based learning for gait recognition using spatio-temporal interest points. *Image and Vision Computing*, 32(12):1117–1126, 2014.

[21] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[22] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013.

[23] M. Lemke, T. Wendorff, B. Mieth, K. Buhl, and M. Linnemann. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *Journal of Psychiatric Research*, 34:277–283, 2000.

[24] H. Li, D. Li, and X. Luo. Bap: Bimodal attribute prediction for zero-shot image categorization. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1013–1016, 2014.

[25] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren. Make the bag disappear: Carrying status-invariant gait-based human age estimation using parallel generative adversarial networks. In *Proc. of the IEEE 10th Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2019)*, pages 1–9, Sep. 2019.

[26] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017.

[27] R. Liao, Y. Makihara, D. Muramatsu, I. Mitsugami, Y. Yagi, K. Yoshiyama, H. Kazui, and M. Takeda. A video-based gait disturbance assessment tool for diagnosing idiopathic normal pressure hydrocephalus. *IEEJ Transactions on Electrical and Electronic Engineering*, 15(3):433–441, Feb. 2020.

[28] R. Liao, S. Yu, W. An, and Y. Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

[29] M. Livne, L. Sigal, N. F. Troje, and D. J. Fleet. Human attributes

from 3d pose tracking. *Computer Vision and Image Understanding*, 116(5):648–660, 2012.

[30] D. López-Fernández, F. J. Madrid-Cuevas, Á. Carmona-Poyato, M. J. Marín-Jiménez, and R. Munoz-Salinas. The ava multi-view dataset for gait recognition. In *International Workshop on Activity Monitoring by Multiple Distributed Sensing*, pages 26–39. Springer, 2014.

[31] J. Lu and Y.-P. Tan. Ordinary preserving manifold analysis for human age estimation. In *IEEE Computer Society and IEEE Biometrics Council Workshop on Biometrics 2010*, pages 1–6, San Francisco, CA, USA, Jun. 2010.

[32] J. Lu, G. Wang, and T. S. Huang. Gait-based gender classification in unconstrained environments. In *Proc of the 21st International Conference on Pattern Recognition*, pages 3284–3287, 2012.

[33] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi. *Gait Recognition: Databases, Representations, and Applications*, pages 1–15. John Wiley & Sons, Inc., 1999.

[34] Y. Makihara, M. Okumura, H. Iwama, and Y. Yagi. Gait-based age estimation using a whole-generation gait database. In *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, pages 1–6, Washington D.C., USA, Oct. 2011.

[35] Y. Makihara, M. Okumura, H. Iwama, and Y. Yagi. Gait-based age estimation using a whole-generation gait database. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2011.

[36] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *Proc. of the 9th European Conference on Computer Vision*, pages 151–163, Graz, Austria, May 2006.

[37] J. Man and B. Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2006.

[38] M. Marin-Jimenez, F. Castro, N. G. F. de la Torre, and R. Medina-Carnicer. Deep multi-task learning for gait-based biometrics. In *Proc. of 2017 IEEE International Conference on Image Processing*, pages 106–110, 2017.

[39] D. Martinho-Corbishley, M. Nixon, and J. Carter. Analysing comparative soft biometrics from crowdsourced annotations. *IET Biometrics*, 5(4):276–283, December 2016.

[40] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter. Super-fine attributes with crowd prototyping. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1486–1500, 2018.

[41] W. Min, S. Mei, L. Liu, Y. Wang, and S. Jiang. Multi-task deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing*, 29:657–669, 2019.

[42] M. P. Murray. Gait as a total pattern of movement: Including a bibliography on gait. *American Journal of Physical Medicine & Rehabilitation*, 46(1):290–333, 1967.

[43] M. S. Nixon, T. N. Tan, and R. Chellappa. *Human Identification Based on Gait*. Int. Series on Biometrics. Springer-Verlag, Dec. 2005.

[44] M. Nordin and A. Saadoon. A survey of gait recognition based on skeleton mode l for human identification. *Research Journal of Applied Sciences, Engineering and Technology*, 2016.

[45] D. Parikh and K. Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011.

[46] B. Qian, X. Wang, N. Cao, Y.-G. Jiang, and I. Davidson. Learning multiple relative attributes with humans in the loop. *IEEE Transactions on Image Processing*, 23(12):5573–5585, 2014.

[47] D. A. Reid, M. S. Nixon, and S. V. Stevenage. Identifying humans using comparative descriptions. 2011.

[48] I. Rida, N. Almaadeed, and S. Almaadeed. Robust gait recognition: a comprehensive survey. *IET Biometrics*, 8(1):14–28, 2019.

[49] A. Sakata, N. Takemura, and Y. Yagi. Gait-based age estimation using multi-stage convolutional neural network. *IPSJ Transactions on Computer Vision and Applications*, 11(1):4, 2019.

[50] R. N. Sandeep, Y. Verma, and C. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3614–3621, 2014.

[51] A. Shehata, Y. Hayashi, Y. Makihara, D. Muramatsu, and Y. Yagi. Does my gait look nice? human perception-based gait relative attribute estimation using dense trajectory analysis. In *Asian Conference on Pattern Recognition*, pages 90–105. Springer, 2019.

[52] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016.

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[54] Y. Souri, E. Noury, and E. Adeli. Deep relative attributes. In *Asian conference on computer vision*, pages 118–133. Springer, 2016.

[55] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017.

[56] M. Tariq and M. A. Shah. Review of model-free gait recognition in biometrie systems. In *2017 23rd International Conference on Automation and Computing (ICAC)*, pages 1–7. IEEE, 2017.

[57] C. Tirkaz, J. Eisenstein, T. M. Sezgin, and B. Yanikoglu. Identifying visual attributes for object recognition from text and taxonomy. *Computer Vision and Image Understanding*, 137:12–23, 2015.

[58] C. Wan, L. Wang, and V. V. Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2018.

[59] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. 2011.

[60] L. Wang, H. Ning, T. Tan, and W. Hu. Fusion of static and dynamic body biometrics for gait recognition. *IEEE Transactions on circuits and systems for video technology*, 14(2):149–158, 2004.

[61] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, dec. 2003.

[62] S. Wang, J. Joo, Y. Wang, and S.-C. Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3111–3118, 2013.

[63] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, pages 155–168. Springer, 2010.

[64] Y. Wang, W. Zhou, Q. Zhang, and H. Li. Visual attribute-augmented three-dimensional convolutional neural network for enhanced human action recognition. *arXiv preprint arXiv:1805.02860*, 2018.

[65] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016.

[66] F. Xiao and Y. Jae Lee. Discovering the spatial extent of relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1458–1466, 2015.

[67] C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, and J. Lu. 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.

[68] C. Yan, B. Zhang, and F. Coenen. Multi-attributes gait identification by convolutional neural networks. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 642–647. IEEE, 2015.

[69] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9):1832–1842, 2016.

[70] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006.

[71] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu. A study on gait-based gender classification. *IEEE Trans. on Image Processing*, 18(8):1905–1910, Aug. 2009.

[72] D. Zhang, Y. Wang, and B. Bhanu. Ethnicity classification based on gait using multi-view fusion. In *IEEE Computer Society and IEEE Biometrics Council Workshop on Biometrics 2010*, pages 1–6, San Francisco, CA, USA, Jun. 2010.

[73] F. Zhang, X. Kong, and Z. Jia. Attribute-correlated local regions for deep relative attributes learning. *Journal of Electronic Imaging*, 27(4):043021, 2018.

[74] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 33–42, 2013.

[75] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu. Robust relative attributes for human action recognition. *Pattern Analysis and Applications*, 18(1):157–171, 2015.

[76] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan. Robust view transformation model for gait recognition. In *2011 18th IEEE International Conference on Image Processing*, pages 2073–2076. IEEE, 2011.