Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Annotator-dependent uncertainty-aware estimation of gait relative attributes

Allam Shehata[a,b,*], Yasushi Makihara[a], Daigo Muramatsu[c], Md Atiqur Rahman Ahad[d], Yasushi Yagi[a]

[a] *Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki 567-0047, Japan*
[b] *Department of Informatics, Electronics Research Institute, Cairo 12622, Egypt*
[c] *Seikei University, Japan*
[d] *Department of Computer Science and Digital Technologies, University of East London, University Way, London E16 2RD, UK*

ARTICLE INFO

Article history:
Received 1 December 2021
Revised 25 October 2022
Accepted 20 November 2022
Available online 22 November 2022

Keywords:
Gait relative attribute
Relative label distribution
Relative score distribution
Annotator's uncertainty
Transition matrix

ABSTRACT

In this paper, we describe an uncertainty-aware estimation framework for gait relative attributes. We specifically design a two-stream network model that takes a pair of gait videos as input. It then outputs a corresponding pair of Gaussian distributions of gait absolute attribute scores and annotator-dependent gait relative attribute label distributions. Moreover, we propose a differentiable annotator-independent uncertainty layer to estimate the gait relative attribute score distribution from the absolute distributions then map it to a relative attribute label distribution using the computation of cumulative distribution functions. Furthermore, we propose another annotator-dependent uncertainty layer to estimate the uncertainty on the gait relative attribute labels in terms of a set of trainable transition matrices. Finally, we design a joint loss function on the relative attribute label distribution to learn the model parameters. Experiments on two gait relative attribute datasets demonstrated the effectiveness of the proposed method against baselines in quantitative and qualitative evaluations.

1. Introduction

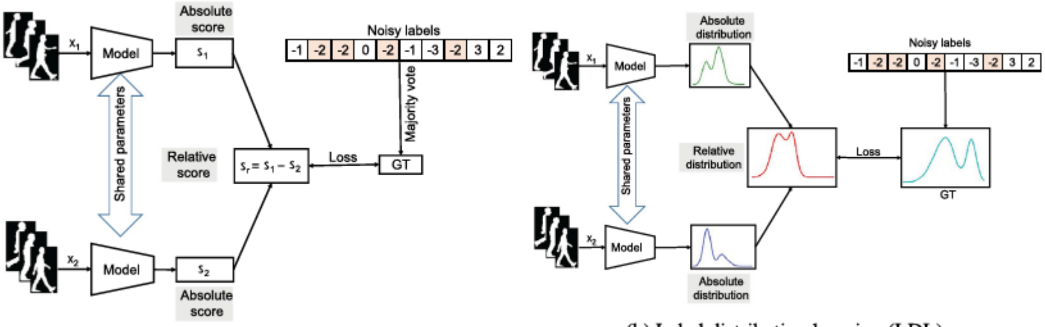
Relative attributes were introduced to serve as a high-level semantic representation of pattern features, which benefits various recognition [1–3] and classification tasks [4–7]. Additionally, one of the core merits of the relative attribute is that it can capture general semantic relationships and enable the relative annotation of instances instead of categorical labels. Therefore, it makes the pattern classification task more beneficial in many applications [8–11]. Despite the ease of annotating in a relative manner, different annotators may assign different scores to the same attribute. This is because of the level of confusion and perception (e.g., individual sense or preference) in the annotation task. Therefore, this may significantly affect the performance of machine learning algorithms (i.e., principally supervised) that build on limited yet imperfect labeled data. Principally, a substantial degree of uncertainty among annotators makes the classification task more challenging. This has led to much interest in developing methods to address the underlying skill levels of human annotators (i.e., uncertainty) associated with labeled data. Recently, treating annotator-uncertainty in the relative attribute estimation task has attracted the pattern recognition and vision community [2,5–7,10]. Understanding various relative attributes (e.g., in the domain of image/video classification, quality assessment, and recognition) while having human annotators and the corresponding uncertainty is a challenging task.

This domain has been much less explored to date, especially by the human gait community. Therefore, in this work, we address the handling of annotator uncertainty in the domain of human gait relative attributes estimation. Generally, recognizing people by their gaits (i.e., walking styles) has increased in popularity recently. Gait is an unobtrusive biometric that can be captured remotely, even from a low-resolution video [12]. Moreover, gait contains not only identity but also various information such as, age [13–15], gender [16–19], emotion [20], ethnicity [21], and human perception-based gait attributes (e.g., goodness and gracefulness) [7,10,22].

Remarkably, human perception-based gait attributes have become more prominent in delineating the human gait [10]. For instance, some people pay attention to their walking style because

[*] Corresponding author at: Department of Informatics, Electronics Research Institute, Cairo 12622, Egypt.
E-mail addresses: allam@am.sanken.osaka-u.ac.jp, allam@eri.sci.eg (A. Shehata).

https://doi.org/10.1016/j.patcog.2022.109197



(a) Score-based learning.

(b) Label distribution learning (LDL).

Fig. 1. Two possible learning frameworks for gait relative attributes using the noisy labels of multiple annotators.

of healthcare or fashion. Therefore, we could develop a smart walking assessment system to advise people about their gait style, and thereby, possible approaches/exercises to improve. Moreover, we could investigate criminals from any video surveillance scene based on their gait attributes (e.g., imposing, nervous walking, large arm swing, wide step length, or curved spine). Therefore, defining the required set of gait attributes is essential to judge a person's walking style, which should be human-understandable and easy to detect. Because of that, in recent studies, researchers have started to address this task by constructing gait attribute datasets and proposing methods for automatic gait attribute estimation [7,10,22–24]. Additionally, it is not necessarily easy to annotate gait attributes (e.g., annotators may not have absolute confidence when annotating the *goodness* of gait), unlike other explicit labels such as identity and age. A possible solution is to obtain the gait attribute annotation in a relative (or comparative) manner [1]. Annotators may find it easy to assess and relate the gait attribute of a person if the input data come in pairs (e.g., the first gait attribute is better than the second gait attribute in the pair). It is more beneficial to have a system that reports the relative scores for a specific gait relative attribute while preserving the underlying uncertainty of the annotation labels instead of reporting only an absolute score (Fig. 1). The existing methods that handled annotators uncertainty (e.g., on image or action quality assessment [25–27], or age estimation [15,24]) considered the estimation of the underlying uncertainty of *absolute labels* – not their *relative labels*. When it comes to gait, the uncertainty-aware approach was introduced in the estimation of gait relative attributes in Makihara et al. [7], in which the label distribution learning (LDL) framework, as shown in Fig. 1(b) was adopted. The authors designed a model that outputs absolute score distributions for the input pair and then estimates the relative score distribution from the pair of the absolute distributions using a trade-off optimal transportation model. Although the method in Makihara et al. [7] can estimate annotator-independent uncertainty, it is difficult to disentangle annotator-independent uncertainty and annotator-dependent uncertainty (e.g., annotator A is more uncertain than annotator B) which is beneficial to understand the performance of the proposed model. Moreover, the method considers estimating the discrete score distribution (absolute or relative) composed of seven bins in a non-parametric manner, which requires more parameters to be estimated and also cannot consider ordinary information explicitly, unlike a continuous parametric distribution, such as the Gaussian distribution. Because of the above-mentioned drawbacks, the method may be unsuitable, particularly in the case in which a relatively small number of annotators (e.g., only 10 annotators) are available. To alleviate the above-mentioned challenges, we propose an annotator-dependent uncertainty-aware model to estimate gait relative attributes. The model can output both annotator-independent and annotator-dependent label distributions of the gait relative attribute by training an annotator-dependent uncertainty model in an end-to-end manner. The contributions of this study are summarized as follows:

1. **An annotator-dependent uncertainty-aware estimation of gait relative attributes:** Unlike the existing uncertainty-aware gait relative attribute estimation framework [7], which does not consider annotator-dependent uncertainty, in this study, we consider both annotator-independent and annotator-dependent uncertainty in a unified framework. Specifically, we adapt the crowd layer used in Rodrigues and Pereira [28] for multi-label image classification to our gait relative attribute estimation problem, and use it to train each annotator's uncertainty (or preference) in an end-to-end manner.
2. **A differentiable conversion module from a continuous score distribution to a discrete label distribution:** A discrete label distribution of gait absolute attributes was used in a previous study [7]; however, we use parametric continuous score distributions of gait absolute/relative attributes, which are suitable for a small number of annotators. Because the annotation is provided in the form of the discrete label of gait relative attributes, our framework should be compatible with it in the training stage. Therefore, we propose a module that converts parametric continuous score distributions into discrete label distribution using a differentiable cumulative distribution function of Gaussian distributions and trainable parameters of interval boundaries for the integral.

2. Related work

Relative attributes An earlier approach to the relative attribute framework was introduced in Parikh and Grauman [1]. The authors introduced the *relative attribute* notion by defining a set of high-level semantic properties in the input instead of absolute labels. In this framework, a pair of training samples are shown to annotators, and they provide relative scores (e.g., the first sample is better or similar, or the second sample is better) based on human perception. Moreover, the authors learned a ranking function for each attribute and then used these learned ranking functions to predict the relative strength of each semantic property in the test input. In [4], the authors extended the concept of relative attributes for ranking and image retrieval based on multi-attribute queries. Compared with existing retrieval approaches that train separate classifiers for each word and ignore inter-dependencies among query terms, this model provides a principled approach for multi-attribute retrieval. It can explicitly model the correlations between attributes. An extension was proposed in Xiao and Jae Lee [2] to discover the relative attributes' spatial extent in the input image pair. Here, a novel formulation was introduced to combine a detector with local smoothness to discover chains of visually coherent patches, generate additional candidate chains, and rank each chain according to its relevance to the attribute.

Inspired by the success of convolutional neural networks, several papers have proposed methods that follow the concept of relative attributes. Papers [6,29] introduced 2-deep relative attribute (DRA) frameworks to learn visual properties from the input images, and use practical 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 unordered image pairs (i.e., similar). Both of these studies were earlier attempts to combine relative attribute estimation with deep learning models in one end-to-end task.

Gait absolute/relative attribute The concept of the visual attribute was used recently in gait analysis to improve performance. In [30], attribute-based classification was applied for gait recognition enhancement by reducing the classifier models required for recognizing each probe gait. This process significantly reduced the computational complexity in the testing phase, in addition to improving the recognition accuracy. The authors of Yan et al. [31] 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 was proposed in Reid et al. [32] based on a set of human soft attributes. An attribute discovery model was proposed in Chen et al. [33] for multi-gait recognition. Stable and discriminative attributes are developed using a latent conditional random field (L-CRF) model that uses the extracted gait energy image to automatically discover the unchanged features from the training images. In the recognition process, the attribute set of each person is detected by inferring on the trained L-CRF model.

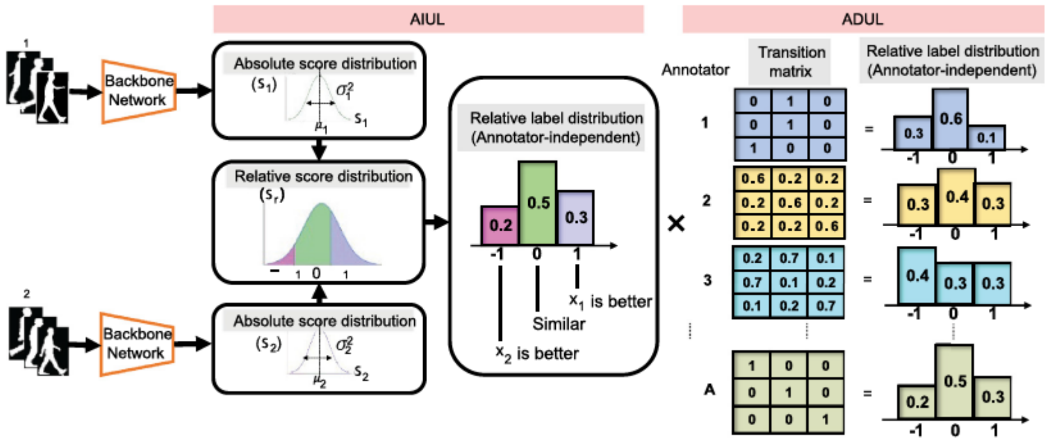
Although it has become easier to use gait attributes (e.g., age and gender) to train models and use them to identify people based on their gait, it is still difficult for such models to recognize the gait who have never been seen before, or relate them to observed people based on their gait attributes [10,22]. Inspired by the relative attributes, a super-fine attribute concept was introduced in Martinho-Corbishley et al. [34] to discover more relevant and precise human descriptions used for person re-identification.

Recently, the concept of the relative attribute was introduced comprehensively in the gait community in Shehata et al. [10]. The authors proposed a motion-based representation using dense trajectories to express walking dynamics. To estimate the relative gait attributes, they trained a set of ranking functions using a Rank Support Vector Machine (Rank-SVM) classifier. Generally, Rank-SVM is used to solve certain ranking problems via learning to rank criteria. These ranking functions estimated a score that indicated each attribute's strength for each walking subject. As an extension, a deep learning-based model was introduced in Hayashi et al. [22] to estimate the gait relative attribute for input gait pairs. The authors also proposed a suitable signed contrastive loss function to train network parameters with the relative annotation. This proposed model achieved better or comparable accuracy for relative attribute prediction compared with the baseline methods. It is worth noting that both of the above-mentioned gait relative attribute methods [10,22] follow the score-based approach shown in Fig. 3 a. In such models, the annotators' uncertainties are completely discarded (i.e., for multi-label datasets, because they consider the majority voting label for model training). Consequently, this makes these approaches less effective in real-world cases. In the proposed approach, we attempt to overcome this by building a relative estimation model that can learn directly from the noisy labels of annotators to obtain the relative label distribution and individual annotator uncertainty in one end-to-end task.

Label distribution learning (LDL) A label distribution learning (LDL) framework is introduced to describe a pattern by a *distribution* rather than the original multiple labels [35]. However, crowdsourcing annotation is being explored as an efficient and cost-effective solution for labeling vast datasets. In fact, the aggregated labels may be assigned by annotators with different levels of expertise or perceptions. This may lead to models with limited predictive performance if they treat these noisy labels as ground-truth labels. Furthermore, in many classification problems, multiple annotation labels may be incorporated into a single visual entity. Compared with single-label learning, this may widen the gap between the model prediction and ground-truth during the training stage. This would result in an inconsistency between the training and test stages [36]. A substantial degree of uncertainty among annotators makes the classification task more challenging [27]. To handle this inconsistency, LDL was exploited in Geng [35] as a generalization learning paradigm of multi-label learning.

Recently, LDL has demonstrated its effectiveness for various computer vision tasks, including age estimation [35,37], pose estimation [38–40], and several other cases [25,41–43]. The authors of Tang et al. [25] proposed an uncertainty-aware score distribution learning method for sports action quality assessment, where the scores were assigned by multiple judges. Moreover, the authors proposed an approach to disentangle the components of the predicted scores using a multi-path uncertainty-aware score distribution learning method. The authors of Mnih and Hinton [44] proposed a method to use noisy labels to model the annotator's uncertainty in terms of a transition matrix and incorporated it into a deep learning model for single/multi-class aerial image classification. Following the same concept, a deep learning framework using noisy labels was proposed in Rodrigues and Pereira [28], where the notion *crowd layer* was introduced to encode the annotator's confusion. The crowd layer demonstrated the ability to directly train deep neural networks from the noisy labels of multiple annotators through only the back-propagation process. As a result, the optimized weight matrices of the crowd layer were introduced to encode the uncertainty of the individual annotators.We inspired from this approach to utilize the annotator uncertainty estimation to serve in the relative attribute task for the first time. Tanno et al. [27] introduced an extension to estimate the annotator uncertainty from noisy label learning. Considering that no actual labels were available, the authors simultaneously learned the annotator confusion and actual underlying distribution. This approach is relevant to our proposed approach in terms of the estimation of annotator confusion (i.e., uncertainty). Although the method proposed in Tanno et al. [27] was used in several applications, such as image classification and medical image assessment, the authors did not propose their method for use in the relative attributes task.

Recently, the gait community started to use the label distribution approach. In [37], the authors proposed an LDL-based framework for age estimation using the gait feature. The proposed framework can model the outputs of discrete label distributions in the absolute age domain. This age label distribution implicitly encodes uncertainty about the estimated age. Despite this, the framework cannot handle relative age distribution estimation, which is common in the age group estimation task (i.e., report if a person is younger, a similar age, or older). However, such system considered only age attribute and did not consider it in a relative manner. A recent trade-off optimal transport model was proposed in Makihara et al. [7] for estimating the gait relative attribute distribution from absolute distributions. Using this model, annotator-independent uncertainty can be treated effectively, whereas annotator-dependent uncertainty is almost dis-



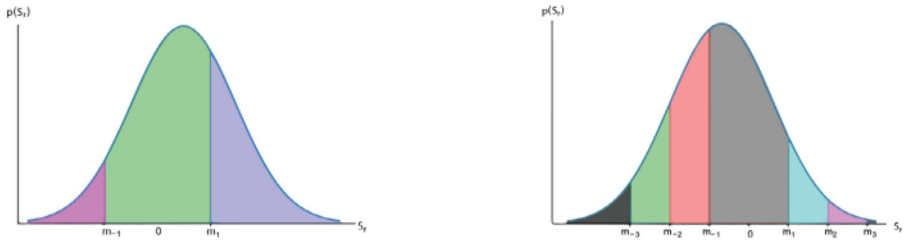
$$1: \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.3 & 0.6 & 0.1 \end{bmatrix}$$

$$2: \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}$$

$$3: \begin{bmatrix} 0.2 & 0.7 & 0.1 \\ 0.7 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.7 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \end{bmatrix}$$

$$A: \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.5 & 0.3 \end{bmatrix}$$

Fig. 2. Proposed learning framework: absolute score distributions are estimated for the input subject pair. Then the annotator-independent uncertainty layer (AIUL) converts the parametric continuous relative score distribution $p(s_r)$ to a discrete relative label distribution $\boldsymbol{p}$ using the trainable parameters of the integral intervals' boundaries $m_{|y|}$ of the comparative labels. The annotator-dependent uncertainty layer (ADUL) proposed to estimate the uncertainty of individual annotators by optimizing a trainable transition matrix of the a-th annotator's T^a to estimate the corresponding annotator label distribution $\boldsymbol{p}^a$.



(a) Three-grade scale.

(b) Seven-grade scale.

Fig. 3. Definition of the lower and upper bounds defined in Eq. (3) for the relative score distributions of the three-grade scale (left) and seven-grade scale (right). Best viewed in color.

carded. Compared to this trade-off model, the proposed model can output both annotator-independent and annotator-dependent uncertainties of the gait relative attribute in an end-to-end manner.

3. Uncertainty-aware gait relative attribute estimation

3.1. Overview

We first provide an overview of the proposed method, which is composed of the annotator-independent uncertainty layer (AIUL) and annotator-dependent uncertainty layer (ADUL), as shown in Fig. 2. In the AIUL, given a pair of gait silhouette sequences, a Siamese network with parameter sharing outputs a pair of parametric continuous absolute score distributions of the gait attribute. We compose our backbone network for a pretrained GaitSet model [45] and a set of fully connected layers following the GaitSet. Then we convert the pair of absolute score distributions into a parametric continuous relative score distribution and then convert it into a discrete relative label distribution with the trainable parameters of the integral intervals' boundaries. In the ADUL, we convert the annotator-independent relative label distribution to annotator-dependent relative label distributions via annotator-specific trainable transition matrices. We describe the proposed model in more detail in the following subsections.

3.2. Problem formulation

First, we provide a brief description of relative labels and the existing relative score learning framework. Then we describe the formulation of the proposed relative LDL framework in detail in the subsequent subsections.

3.2.1. Relative labels

In the relative attribute framework, instead of annotating absolute attribute scores (or categorical labels) for each training sample, an annotator is given a pair of training samples and assigns a relative attribute score (or categorical labels) for the pair. Specifically, the annotator may assign a positive value when the first sample is better than the second sample in the pair, and vice versa. Consequently, a zero value naturally means that both samples are similar. It is worth noting that the absolute value assigned by the annotator indicates the degree of "goodness" of the desired attribute. For instance, an annotator may assign +3 when the first sample is more beautiful than the second sample, whereas the annotator may assign +1 when the first sample is slightly better than the second. Generally, we define the relative label set $\mathcal{Y} = \{-M, \dots, -1, 0, 1, \dots, M\}$, where M depends on the dataset annotation. For example, $M = 1$ for the 3-grade scale annotation proposed in Shehata et al. [10], whereas $M = 3$ for the 7-grade scale annotation introduced in Makihara et al. [7].

3.2.2. Relative score estimation

Given a gait silhouette sequence x, we aim at estimating an absolute score s of a certain gait attribute as $s = f(x)$, where $f(\cdot)$ is a mapping function from the gait silhouette sequence to the gait absolute attribute score. We generally implement the mapping function f using a deep neural network, similar to Yang et al. [6], Hayashi et al. [22], Souri et al. [29]. Once we collect N pairs of gait silhouette sequences and their corresponding relative labels as $\{x_{1,i}, x_{2,i}, y_i\} (i = 1, \dots, N)$, we train the mapping function f so that the estimated relative score $s_{r,i}$ (i.e., difference between the absolute scores) is

$$s_{r,i} = s_{1,i} - s_{2,i} = f(x_{1,i}) - f(x_{2,i}). \tag{1}$$

This relative score is consistent with the relative label (e.g., if a relative label y is positive, we expect the estimated relative score to be positive, and hence, $f(x_1) - f(x_2) > 0$).

When multiple relative labels are assigned by multiple annotators for the same pair, we may consider the majority voting label of relative noisy annotations as the ground-truth label to train the mapping functions. However, the main drawback of the majority voting label is that it cannot consider annotator-dependent uncertainty derived from, for example, different skills and preferences.

3.2.3. Relative label distribution estimation

Unlike the above-mentioned relative score estimation, which does not consider uncertainty, we aim at estimating a relative label distribution, which considers uncertainty. We also use a parametric continuous absolute score distribution that requires fewer parameters than the existing method [7]. Specifically, we select the Gaussian distribution, which is defined by only two parameters, that is, mean and variance, and regard the variance as the degree of uncertainty. Let a distribution of an absolute score s_1 for the first input gait x_1 be $\mathcal{N}(s_1; \mu_1, \sigma_1^2)$, where μ_1 and σ_1^2 are the predicted mean μ_1 and variance σ_1^2, respectively. Similarly, we introduce a distribution for the second input gait x_2 as $\mathcal{N}(s_2; \mu_2, \sigma_2^2)$. We then define a distribution of a relative score s_r $(= s_1 - s_2)$ as a Gaussian distribution $\mathcal{N}(s_r; \mu_r, \sigma_r^2)$ with mean μ_r and variance σ_r^2, which are computed based on the *i.i.d* assumption $\mu_r = \mu_1 - \mu_2$ and $\sigma_r^2 = \sigma_1^2 + \sigma_2^2$. Consequently, the specific form of the probability distribution function for the relative score s_r is

$$p(s_r) = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left\{-\frac{(s_r - \mu_r)^2}{2\sigma_r^2}\right\}. \tag{2}$$

Because the annotation is assigned in the form of a discrete relative label, we need to convert the continuous score distribution $p(s_r)$ into the discrete label distribution $P(y)$, where $y \in \mathcal{Y}$ is a relative label.

To achieve this, we define an interval of the relative score, which the relative label y occupies, and then compute the corresponding probability by integral. Specifically, we define the symmetric boundary parameters $\{m_{|y|}\}(|y| = 1, \ldots, M)$ of the intervals, as shown in Fig. 3, and then define the lower bound $l(y)$ and upper bound $u(y)$ of the interval for the relative label y as

$$l(y) = \begin{cases} -\infty & (y = -M) \\ -m_{|y|+1} & (y \leq 0) \\ m_y & (y > 0) \end{cases}, \quad u(y) = \begin{cases} -m_{|y|} & (y < 0) \\ m_{y+1} & (y \geq 0) \\ \infty & (y = M) \end{cases} \tag{3}$$

We then compute the probability of the relative label y:

$$P(y) = \int_{l(y)}^{u(y)} p(s_r) ds_r. \tag{4}$$

Note that if we use a three-grade relative label, as in Shehata et al. [10], the number of boundary parameters is one (i.e., m_1), whereas it is three (i.e., m_1, m_2, m_3) for a seven-grade relative label, as in Makihara et al. [7]. Additionally, note that the boundary parameters are trainable to be more consistent with the annotated relative labels; that is, they are not hyperparameters. Moreover, note that we use a fixed value $m_1 = 1$ in the case of a single boundary (i.e., three-grade case with $M = 1$) because it is equivalent to moving boundary m_1 while changing the scale of the score, and we change the scale of the score by changing the backbone network parameters (i.e., it is redundant to use both). Finally, we compute the probabilities for all the relative labels $y \in \mathcal{Y}$ and then define a probability vector whose entities are the probability as $\boldsymbol{p} = [P(-M), \ldots, P(M)] \in \mathbb{R}^{1\times(2M+1)}$. Note that the probability vector $\boldsymbol{p}$ describes annotator-independent (i.e., global) uncertainty.

3.2.4. Annotator-dependent uncertainty-aware estimation

In this subsection, we describe a method to convert the annotator-independent relative label distribution to the annotator-dependent relative label distribution. Assuming that we have A annotators, we define the relative label distribution $\boldsymbol{p}^a = [P^a(-M), \ldots, P^a(M)] \in \mathbb{R}^{2M+1}$ for the ath annotator. Specifically, we assume that the annotator-dependent label distribution $\boldsymbol{p}^a$ of the ath annotator is represented by a linear transformation of the annotator-independent relative label distribution $\boldsymbol{p}$:

$$\boldsymbol{p}^a = \boldsymbol{p}T^a, \tag{5}$$

where $T^a \in \mathbb{R}^{(2M+1)\times(2M+1)}$ is a transition matrix for the ath annotator. Note that the (i, j)th component $t_{i,j}^a$ of the transition matrix T^a for the ath annotator indicates a conditional probability of the ath annotator's relative label $y^a = j$ given the annotator-independent relative label $y = i$, i.e., $t_{i,j}^a = p(y^a = j|y = i)$.

We consider the transition matrix to be stochastic, where its entries are non-negative real numbers that represent a probability, and the summation over columns for each row is 1. To enforce this property using the unconstrained trainable parameters of a deep neural network, we use a softmax function. Specifically, we introduce an unconstrained parameter π_{ij}^a, which corresponds to the (i, j)th component of matrix T^a. So, the normalized transition probability $t_{i,j}^a$ is defined as

$$t_{ij} = \frac{\exp(\pi_{ij})}{\sum_j \exp(\pi_{ij})}. \tag{6}$$

Then we optimize the unconstrained transition parameters $\{\pi_{i,j}^a\}$ through training. Note that the optimized transition parameters are considered to implicitly encode the individual annotator's uncertainty and well describe the annotator's skill levels about the relative attribute annotation.

3.3. Loss function

We introduce two loss functions for annotator-dependent relative labels and annotator-independent relative labels. First, we denote the ground-truth relative label y_i^{a*} given by the ath annotator for the ith training sample, and a corresponding relative label distribution $\boldsymbol{p}_i^a = [P_i^a(-M), \ldots, P_i^a(M)] \in \mathbb{R}^{2M+1}$ estimated using our model. Similar to other general label-based classification tasks, we use the cross-entropy loss and sum over the annotators and the training samples as,

$$L_{\mathrm{CE}} = \sum_{i=1}^{N} \sum_{a=1}^{A} \sum_{y=-M}^{M} -\delta_{y,y_i^{a*}} \log P_i^a(y), \tag{7}$$

where N is the number of training samples and δ is the Kronecker delta. Additionally, we represent the ground truth of the annotator-dependent relative label distribution as a one-hot encoding distribution. We also compute the ground-truth annotator-independent relative label distribution $\boldsymbol{p}_i^* = [P_i^*(-M), \ldots, P_i^*(M)] \in \mathbb{R}^{1\times 2M+1}$ for the ith training sample by simply aggregating all the annotators' relative labels as follows:

$$P_i^*(y) = \frac{1}{A} \sum_{a=1}^{A} \delta_{y,y_i^{a*}}. \tag{8}$$

Because the ground-truth annotator-independent relative label distribution $\boldsymbol{p}_i^*$ is not necessarily a one-hot distribution, unlike the ground-truth annotator-dependent relative label distribution, we use a loss function suited for LDL. Specifically, given the ith sample, we compute the Jensen–Shannon divergence (JSD) between the ground-truth annotator-independent relative label distribution $\boldsymbol{p}_i^*$ and an estimated annotator-independent relative label distribution $\boldsymbol{p}_i = [P_i(-M), \ldots, P_i(M)] \in \mathbb{R}^{2M+1}$ and sum over all the training samples as follows:

$$L_{\mathrm{JS}} = \sum_{i=1}^{N} D_{\mathrm{JS}}(\boldsymbol{p}_i^* \parallel \boldsymbol{p}_i) \tag{9}$$

$$= \sum_{i=1}^{N} \left\{ \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{p}_i^* \parallel \bar{\boldsymbol{p}}_i) + \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{p}_i \parallel \bar{\boldsymbol{p}}_i) \right\}, \tag{10}$$

where $\bar{\boldsymbol{p}}_i = [\bar{P}_i(-M), \dots, \bar{P}_i(M)] \in \mathbb{R}^{2M+1}$ is the mean distribution of the two distributions $\boldsymbol{p}_i^*$ and $\boldsymbol{p}_i$ (i.e., $\bar{\boldsymbol{p}}_i = (\boldsymbol{p}_i^* + \boldsymbol{p}_i)/2$), and $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ is the Kullback–Leibler (KL) divergence. For example, the KL divergence $D_{\mathrm{KL}}(\boldsymbol{p}_i^* \parallel \bar{\boldsymbol{p}}_i)$ is computed as

$$D_{\mathrm{KL}}(\boldsymbol{p}_i^* \parallel \bar{\boldsymbol{p}}_i) = \sum_{y=-M}^{y=M} P_i^*(y) \log \left(\frac{P_i^*(y)}{\bar{P}_i(y)} \right), \tag{11}$$

and the KL divergence $D_{\mathrm{KL}}(\boldsymbol{p}_i \parallel \bar{\boldsymbol{p}}_i)$ is computed in a similar manner. Finally, the loss function is the summation of the above two loss functions:

$$L = L_{\mathrm{CE}} + L_{\mathrm{JS}}. \tag{12}$$

To summarize, given a set of training data $\{x_{1,i}, x_{2,i}, y_i^{a*}\}_{i=1:N}^{a=1:A}$, we optimize the set of intervals' boundaries $\{m_k\}_{k=1}^{M}$, transition matrices $\{T^a\}_{a=1}^{A}$, and the backbone network parameters by minimizing the loss function L.

4. Evaluation experiments

4.1. Dataset

4.1.1. Three-grade dataset [10]

To the best of our knowledge, the first gait relative attribute dataset was introduced in Shehata et al. [10], Hayashi et al. [22]. This dataset was compiled from the publicly available gait recognition dataset, OULP-Age [23]. A set of 1200 subjects' walking videos were compiled from this dataset, then arranged into pairs of subjects, and presented to several annotators for comparative annotation. Additionally, eight gait attributes have been defined: {*General goodness, Stately, Cool, Relax, Arm swing, Walking speed, Step length, Spine*}. Each attribute describes a certain visual property of the walking subject ranging from perceptual attributes (e.g., relaxed vs. nervous, and happy vs. sad) to physical attributes (e.g., step length and arm swing). Each attribute could receive comparative labels from the ternary set $\mathcal{Y} = \{1, 0, -1\}$.

Because of the limited number of annotated pairs for this dataset, we adopted cross-validation to evaluate the model performance on different folds. Hence, we first split the 1200 walking subjects into 200 for testing and 1000 for training. We repeated this splitting for six folds, and on each fold, we generated different subject pairs for training and testing. Then, for each attribute, we used the 1000 training samples to selectively generate the training pairs and the 200 samples to generate the testing pairs. To completely disjoint the training and testing pairs, we excluded the pairs where a certain subject appeared in both the training and testing pairs. Hence, we had 800 training pairs and 100 test pairs for each attribute.

4.1.2. Seven-grade dataset [7]

For this dataset, we chose the walking videos of 1200 subjects pairs from the largest multi-view dataset, OUMVLP [46]. The authors in Makihara et al. [7] hired ten annotators and designed an annotation tool as shown in Fig. 5. The annotators watched gait silhouette sequences from the side and front views, and assigned annotation labels for five gait attributes: *beautiful, cheerful, imposing, relax*, and *graceful*. Each annotator was asked to select one of the seven grades for each attribute. The grades had values from 3 (leftmost) to −3 (rightmost): grades 3, 2, and 1 trivially indicated that the first sample was much better, better, or slightly better, respectively; grade 0 was neutral (i.e., the attributes were similar); and grades −1, −2, and −3 indicated that the second sample was slightly better, better, or much better, respectively. Finally, we arranged 1080 pairs for training and 120 pairs for testing. For training the score-based baselines, we squeezed the seven grades into three grades to meet the loss function requirements of those baselines. However, we kept the seven-grade setting for the testing stage.

4.1.3. Annotation statistics

Fig. 4 shows the diversity of uncertainty among the annotators for both datasets. Specifically, we first computed the average and standard deviation (SD) of the grades over six (three classes dataset [10]) and 10 (seven classes dataset [7]) annotations for each pair and each attribute. We then computed the histograms of the average and SD over the entire set of subject pairs for each attribute. Fig. 4 shows that the averaged grades for both datasets were almost distributed symmetrically around zero. The SD for the three-grade scale annotation [10] was distributed mainly between 0.15 and 1, whereas the SD for the seven-grade scale annotation [7] was distributed between 0.5 and 2.0. Both SD ranges indicate inconsistency (i.e., different skill levels) among the annotators as a result of human perceptions.

4.2. Implementation details

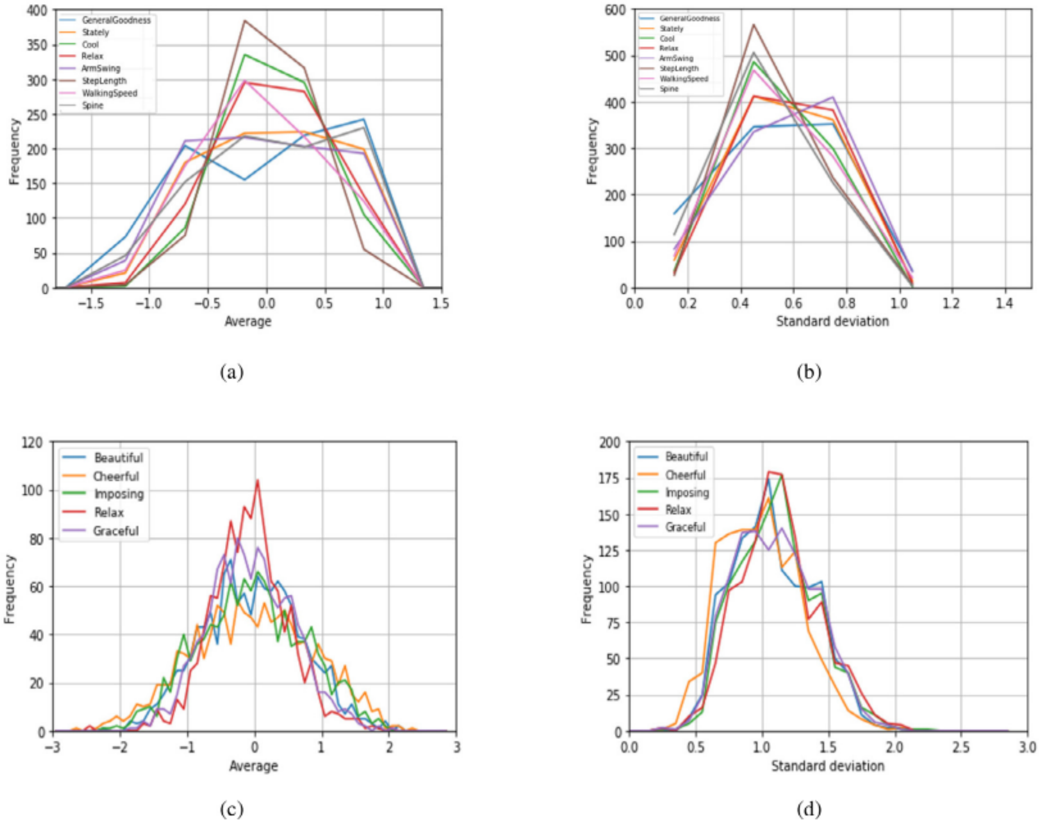
For the three-grade scale dataset [10] experiments, we set the initial learning rate for the fully connected layers to 5×10^{-5} and weight decay to 5×10^{-10} to train the network from scratch. For the ADUL, we set the learning rate to 1×10^{-3} and set the margin parameter of the AIUL to $m = 1$. Additionally, we initialized the transition matrices of the ADUL using the identity matrix. For parameter optimization, we performed mini-batch Adam optimization [47] on the cross-entropy loss. We trained the model to minimize the cross-entropy loss (Eq. (7)), with a batch size of 64 pairs (i.e., 128 silhouette sequences) and 100 epochs.

For the seven-grade scale annotation dataset [7] experiments, we set the initial learning rate of the fully connected layers by rate to 1×10^{-4} and weight decay to 5×10^{-6} to train the network from scratch. Additionally, the learning rates of the AIUL and ADUL layers were 5×10^{-3} and 1×10^{-4}, respectively. For parameter optimization, we performed mini-batch Adam optimization [47] on the combined loss. Also, we initialized the transition matrices of the ADUL using the identity matrix, and the trainable margins of the AIUL $\{m_1, m_2, m_3\}$ to $\{0.5, 1.5, 3\}$, respectively. We trained the model to minimize the combined loss (Eq. (12)), with a batch size of 64 pairs (i.e., 128 silhouette sequences) and 200 epochs.

4.3. Benchmarks

For the performance evaluation, we compared our proposed method with existing score-based and distribution-based baselines. For the score-based methods, the output of the model was a relative score, and we used the majority voting label for training. These baselines do not consider uncertainty or learning transition matrices from noisy annotations. Moreover, for the estimation of the annotator label distribution, a separate model was required for each annotator's annotation, which raises a concern about the storage size and time consumption.

By contrast, the distribution-based baseline assumes that the output of the model is a relative score distribution instead of a single-value relative score. Hence, it is applicable for handling the



(a) (b)

(c) (d)

Fig. 4. Histograms of the average and standard deviation of noisy labels over the annotators. (a),(b) Histograms of the average and standard deviation of the three-grade scale annotation dataset [10], respectively. (c), (d) Histograms of the average and standard deviation of the seven-grade scale annotation dataset [7], respectively.

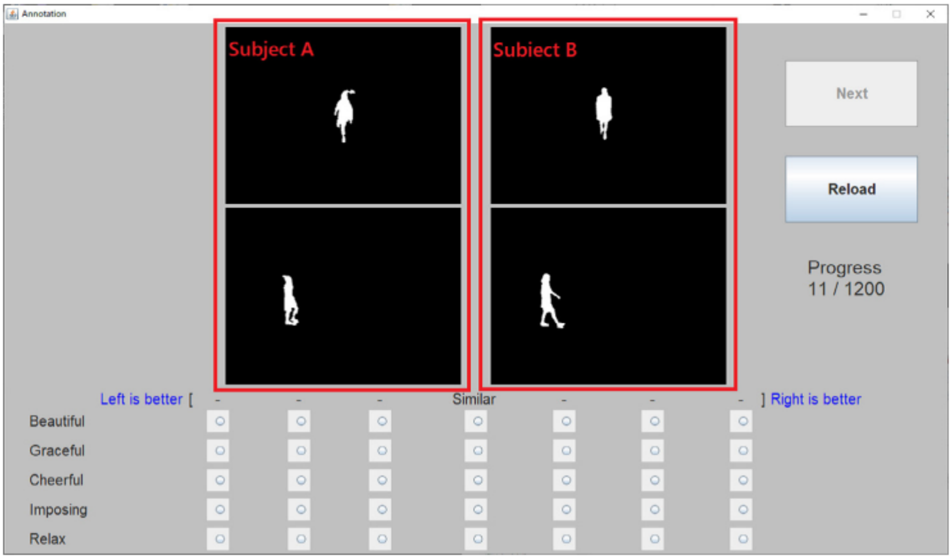


Fig. 5. A screenshot of the seven-grade annotation tool for the five gait relative attributes as proposed in Makihara et al. [7].

annotator's uncertainty and estimating the annotator label distribution from noisy labels in an end-to-end task, which alleviates the drawbacks of score-based baselines. We briefly describe the baselines as follows:

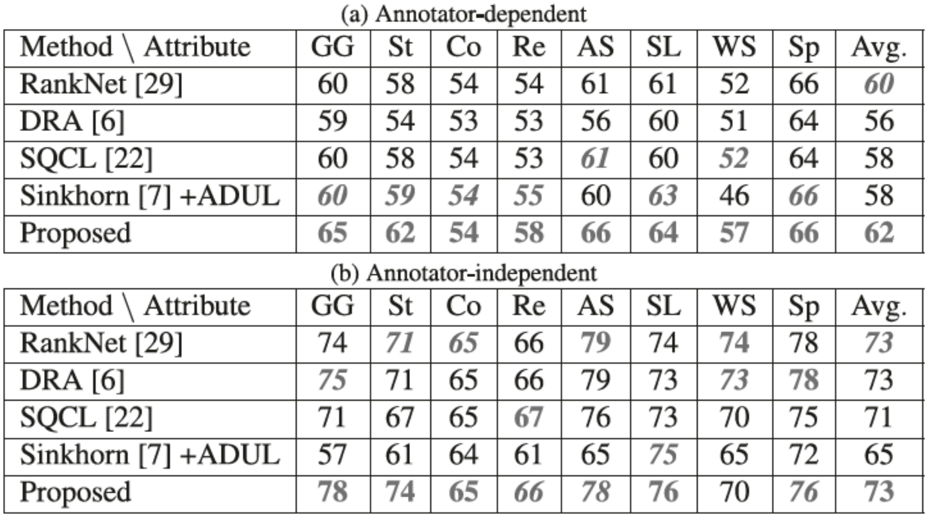
RankNet [29]: This baseline is a deep relative attribute model that was proposed for image classification [29]. The authors proposed using the binary cross-entropy loss for model parameter optimization.

DRA [6]: This baseline is a deep relative attribute model that was proposed for image classification [6]. The authors proposed the signed linear contrastive loss for model parameter optimization.

SQCL [22]: This baseline is a deep relative attribute model used for gait relative attribute estimation [22]. The authors proposed the signed quadratic contrastive loss for model parameter optimization.

Sinkhorn + ADUL [7]: This baseline uses the trade-off optimal transport model to estimate the relative distribution from the absolute distributions [7]. We added the ADUL to estimate both the annotator's uncertainty and label distributions for a fair compar-

Table 1
Classification accuracy [%] ↑ of the relative label with the three-grade dataset. Abbreviations for attributes are general goodness (GG), stately (St), cool (Co), relax (Re), arm swing (AS), step length (SL), walking speed (WS), spine (Sp), and average (Avg.). Bold red and italic bold blue indicate the best and second-best accuracies, respectively. This convention is consistent throughout the paper.



(a) Annotator-dependent

Method \ Attribute	GG	St	Co	Re	AS	SL	WS	Sp	Avg.
RankNet [29]	60	58	54	54	61	61	52	66	***60***
DRA [6]	59	54	53	53	56	60	51	64	56
SQCL [22]	60	58	54	53	***61***	60	***52***	64	58
Sinkhorn [7] +ADUL	***60***	***59***	***54***	***55***	60	***63***	46	***66***	58
Proposed	**65**	**62**	**54**	**58**	**66**	**64**	**57**	**66**	**62**

(b) Annotator-independent

Method \ Attribute	GG	St	Co	Re	AS	SL	WS	Sp	Avg.
RankNet [29]	74	***71***	***65***	66	**79**	74	**74**	78	***73***
DRA [6]	***75***	71	65	66	79	73	***73***	**78**	73
SQCL [22]	71	67	65	**67**	76	73	70	75	71
Sinkhorn [7] +ADUL	57	61	64	61	65	***75***	65	72	65
Proposed	**78**	**74**	**65**	***66***	***78***	**76**	70	***76***	**73**

ison with the proposed method. It is possible distribution-based baseline to compare with for gait relative attribute.

4.4. Evaluation criteria

To evaluate the accuracy of the proposed method against the baselines, we considered the annotation of the test data, in addition to the output of the baselines and proposed method. For the score-based baselines, the test data annotation was given in the form of comparative labels. Therefore, we also evaluated accuracy in a pairwise manner. Furthermore, the last output for each test pair was given as score difference d. As a result, we classified the data into three classes or seven classes using thresholds trained using a greedy search algorithm [10,22]. Note that we selected the thresholds to maximize the classification accuracy of the training data.

The outputs of both the proposed method and Sinkhorn baseline [7] were probability distributions with $(2M+1)$ bins instead of scalar scores. To evaluate the classification accuracy, we chose the class label with the highest probability and evaluated it against the ground-truth label. The ground-truth label was a noisy label of each annotator for the annotator label distribution evaluation (i.e., ADUL), and the majority voting label for the relative label distribution evaluation (i.e., AIUL). Furthermore, we evaluated the dissimilarity of the estimated relative label distribution. We used the JSD (Eq. (10)) to measure the dissimilarity between the estimated relative label distribution and the ground-truth distribution. To obtain the JSD dissimilarity bounded by [0,1], we replaced the natural logarithm in Eq. (10) by the base-2 logarithm. To evaluate the transition matrix evaluation, we used the Frobenius norm [48] to measure the error between the estimated and ground-truth transition matrices.

4.5. Quantitative evaluation

In our quantitative evaluation, we compared the proposed approach with the possible baselines. For all baselines, we adopted the pretrained GaitSet model [45] as the backbone network. We reported the classification accuracy, in addition to, the transition matrix estimation error and JSD dissimilarity.

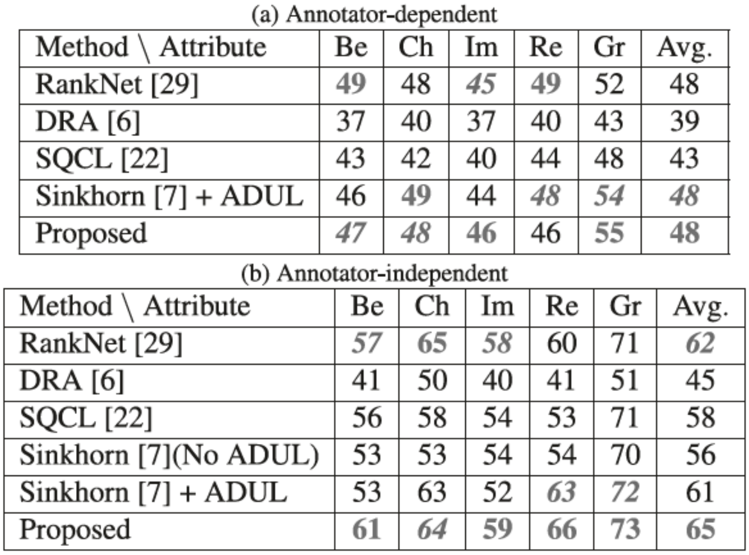
4.5.1. Three-grade dataset

As shown in Table 1(a) and (b), we reported the classification accuracy of the predicted annotator label distribution and relative label distribution, respectively. We conducted this evaluation on the three-grade scale annotation dataset [10,22]. For both tables, the first three rows are for score-based approaches [6,22,29] and the fourth row shows the classification accuracy of the Sinkhorn-based baseline [7] + ADUL.

The proposed approach outperformed the Sinkhorn-based baseline [7], with an average accuracy of 4% for the annotator label distribution and 8% for the relative label distribution estimation. Additionally, its performance was better than or comparable with that of the score-based methods (the first three rows). It is worth noting that the score-based methods use the majority voting label directly for model training and do not consider either learning from noisy labels or handling the annotator's uncertainty.

For instance, although the RankNet approach [29] is principally considered as a scored-based approach, it achieved better or comparable performance. Specifically, the RankNet model received the absolute scores of the input pair, computed the corresponding relative score (i.e., score difference), and then mapped it onto the probability using a logistic function to meet the binary cross-entropy loss. Therefore, RankNet could not produce the absolute distribution of the input pair, which made it difficult to fit and handle the underlying annotation uncertainty. We explain the marginal performance as follows: we trained the RankNet model directly using the comparative label for annotator-dependent attribute score estimation and the majority voting label for annotator-independent relative attribute score estimation. By contrast, we trained the proposed approach using only the annotator-specific noisy labels to estimate the relative label distribution (i.e., global uncertainty), annotators' uncertainties, and annotator-specific label distribution in an end-to-end task. Furthermore, the proposed method learned the annotator label distribution and the annotator's uncertainty using single-model training. By contrast, the score-based methods required a separate model for each individual annotation. Overall, the proposed model was the best or second best compared with the state-of-the-art methods.

Table 2
Classification accuracy [%] ↑ of the predicted relative label with the seven-grade scale dataset. Abbreviations for the attributes: mean beautifulness (Be), cheerfulness (Ch), imposingness (Im), relaxedness (Re), gracefulness (Gr), and average (Avg.).



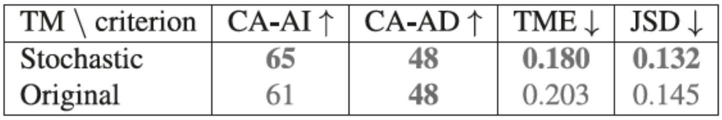
(a) Annotator-dependent

Method \ Attribute	Be	Ch	Im	Re	Gr	Avg.
RankNet [29]	**49**	48	***45***	**49**	52	48
DRA [6]	37	40	37	40	43	39
SQCL [22]	43	42	40	44	48	43
Sinkhorn [7] + ADUL	46	**49**	44	***48***	***54***	***48***
Proposed	***47***	***48***	**46**	46	**55**	**48**

(b) Annotator-independent

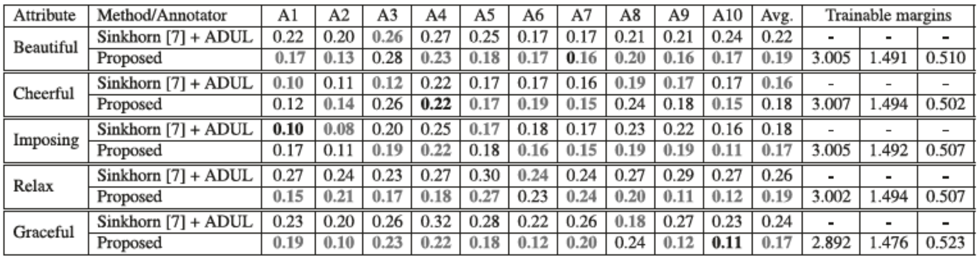
Method \ Attribute	Be	Ch	Im	Re	Gr	Avg.
RankNet [29]	***57***	**65**	***58***	60	71	***62***
DRA [6]	41	50	40	41	51	45
SQCL [22]	56	58	54	53	71	58
Sinkhorn [7](No ADUL)	53	53	54	54	70	56
Sinkhorn [7] + ADUL	53	63	52	***63***	***72***	61
Proposed	**61**	***64***	**59**	**66**	**73**	**65**

Table 3
Effect of different settings for the transition matrix, that is, the stochastic matrix versus the original matrix on criteria with the seven-grade dataset. We report the following criteria averaged over the attributes and annotators accordingly: classification accuracy of the annotator-independent relative label (CA-AI) [%], that of the annotator-dependent relative label (CA-AD) [%], the transition matrix error (TME), and the Jensen–Shannon divergence (JSD). Best performance is marked in bold.



TM \ criterion	CA-AI ↑	CA-AD ↑	TME ↓	JSD ↓
Stochastic	**65**	**48**	**0.180**	**0.132**
Original	61	**48**	0.203	0.145

Table 4
Transition matrix error (Frobenius norm) between the estimated transition matrices (annotator's uncertainty) and the ground-truth matrices for 10 annotators. We considered the seven-grade scale annotation dataset for this evaluation [7]. Lower is better (↓ bold). Last three columns contain the optimized margins $\{m_3, m_2, m_1\}$ used for relative label distribution conversion.



Attribute	Method/Annotator	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Avg.	Trainable margins		
Beautiful	Sinkhorn [7] + ADUL	0.22	0.20	**0.26**	0.27	0.25	0.17	0.17	0.21	0.21	0.24	0.22	-	-	-
	Proposed	**0.17**	**0.13**	0.28	**0.23**	**0.18**	**0.17**	**0.16**	**0.20**	**0.16**	**0.17**	**0.19**	3.005	1.491	0.510
Cheerful	Sinkhorn [7] + ADUL	**0.10**	0.11	**0.12**	0.22	0.17	0.17	0.16	**0.19**	**0.17**	0.17	**0.16**	-	-	-
	Proposed	0.12	**0.14**	0.26	**0.22**	**0.17**	**0.19**	**0.15**	0.24	0.18	**0.15**	0.18	3.007	1.494	0.502
Imposing	Sinkhorn [7] + ADUL	**0.10**	**0.08**	0.20	0.25	**0.17**	0.18	0.17	0.23	0.22	0.16	0.18	-	-	-
	Proposed	0.17	0.11	**0.19**	**0.22**	0.18	**0.16**	**0.15**	**0.19**	**0.19**	**0.11**	**0.17**	3.005	1.492	0.507
Relax	Sinkhorn [7] + ADUL	0.27	0.24	0.23	0.27	0.30	**0.24**	0.24	0.27	0.29	0.27	0.26	-	-	-
	Proposed	**0.15**	**0.21**	**0.17**	**0.18**	**0.27**	0.23	**0.24**	**0.20**	**0.11**	**0.12**	**0.19**	3.002	1.494	0.507
Graceful	Sinkhorn [7] + ADUL	0.23	0.20	0.26	0.32	0.28	0.22	0.26	**0.18**	0.27	0.23	0.24	-	-	-
	Proposed	**0.19**	**0.10**	**0.23**	**0.22**	**0.18**	**0.12**	**0.20**	0.24	**0.12**	**0.11**	**0.17**	2.892	1.476	0.523

4.5.2. Seven-grade dataset

Table 2 (a) shows the classification accuracy of the estimated annotator label distribution of the proposed method compared with the baselines. The proposed method outperformed the DRA [6] and SQCL [22] baselines, whereas its performance was comparable with that of the RankNet [29] and Sinkhorn-based [7] baselines. Similarly, Table 2(b) shows the classification accuracy of the predicted relative label distribution. The proposed method outperformed the baseline methods, with an accuracy improvement ranging from 3% for RankNet [29] model to 20% for the DRA [6] method.

Furthermore, we evaluated the estimation of the annotator's uncertainty. Table 4 shows the transition matrix error between the estimated transition matrix and the ground-truth. The proposed method achieved a better estimation than the Sinkhorn-based baseline [7]. Because the score-based baselines are not applicable for the estimation of an individual annotation and the transition matrix estimation, there is no transition matrix error evaluation for those baselines.

To evaluate the dissimilarity of the predicted relative label distribution, we report the Jensen–Shannon dissimilarity for both the three-grade scales and seven-grade scale annotation datasets in Table 5(a) and (b), respectively. The proposed method with AIUL performed better, which emerges the contribution of the proposed layer for relative label distribution estimation.

Table 5
Jensen–Shannon divergence between the ground-truth and estimated relative label distribution. Lower is better (↓ bold).

(a) Three-grade dataset

Method \ Attribute	GE	St	Co	Re	AS	SL	WS	Sp	Avg.
Sinkhorn [7] + ADUL	0.247	0.231	**0.194**	0.206	0.234	0.178	**0.196**	0.186	0.209
Proposed	**0.140**	**0.160**	0.204	**0.197**	**0.175**	**0.163**	0.202	**0.148**	**0.174**

(b) Seven-grade dataset

Method \ Attribute	Be	Ch	Im	Re	Gr	Avg.
Sinkhorn [7] (No ADUL)	0.306	0.271	0.320	0.276	0.242	0.284
Sinkhorn [7] + ADUL	0.281	0.241	0.244	0.270	0.234	0.254
Proposed	**0.135**	**0.125**	**0.128**	**0.144**	**0.130**	**0.132**

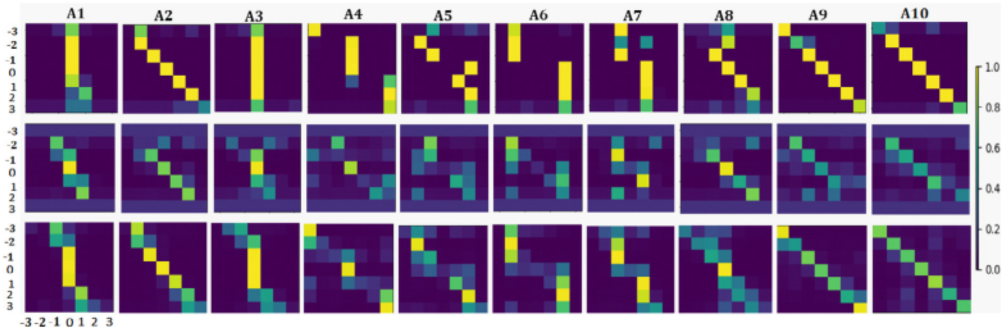


Fig. 6. Qualitative evaluation on the seven-grade scale annotation dataset [7]. We show the estimated transition matrices of 10 annotators versus the corresponding ground-truth for the *Graceful* gait attribute. First row shows the ground-truth, second row shows the prediction of the Sinkhorn baseline [7] + ADUL, and last row shows the proposed method estimation. Transition matrix's rows are normalized so that each row sums to one. Best viewed in color.

4.5.3. Transition matrix evaluation

Additionally, we evaluated the accuracy of the predicted transition matrices in Fig. 6 for the *Graceful* gait attribute. We observed visually that the estimated transition matrices using the proposed method (bottom row) were close to the ground-truth matrices (top row).

4.6. Sensitivity analysis

The key trainable parameters of the proposed model were in the transition matrix. It should be initialized carefully to achieve fast convergence. In this experiment, we used the seven-grade scale annotation dataset to analyze the sensitivity of the transition matrix on the classification accuracy of the proposed method. We reported the accuracy of the predicted distributions, in addition to the distribution dissimilarity under two settings of the transition matrix: the stochastic matrix versus the optimized real-valued matrix. Table 3 shows that the estimation accuracy degraded when we used the real-valued transition matrix compared with using the stochastic matrix, which supports our proposed use of the stochastic transition matrix.

4.7. Discussion

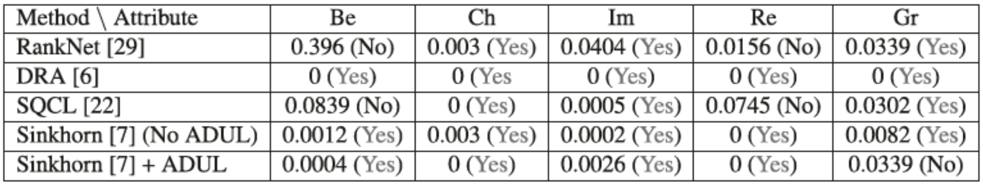
4.7.1. Statistical analysis

We conducted a further experiment to justify the performance significance of our method against the baselines. For each input sample, we estimated the relative scores for the score-based methods and the expectation of the relative distribution for the distribution-based methods. We then computed the absolute errors between the predicted relative scores/expectation and the expectation of the ground-truth distributions for all input pairs. We considered using the statistical test $t_{score} = \frac{\bar{d}}{\sqrt{\frac{s^2}{N}}}$. It is a function of both the mean $\bar{d}$ and variance s^2 of the computed difference between the absolute error pairs of the proposed method and a desired baseline, and N is the total number of input pairs. Essentially, for a higher t_{score} value, a significant difference existed between the proposed method and the baseline. By contrast, the smaller the t_{score}, the more similarity existed and hence, no significant difference existed.

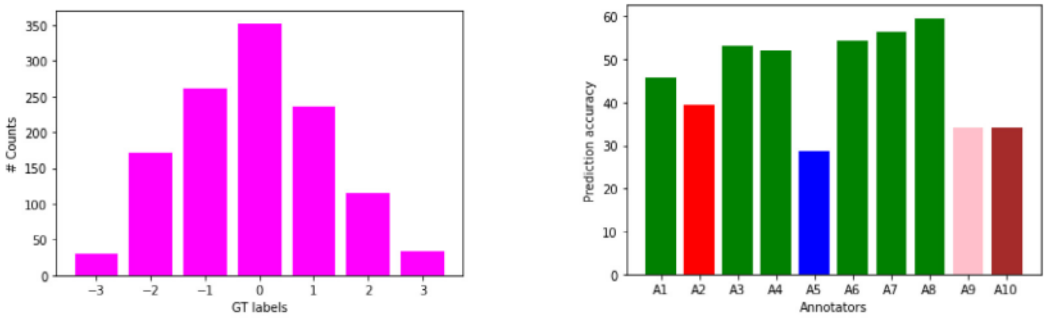
To apply the test, we defined the null hypothesis H_0 as follows: the difference's mean $\bar{d} \geq 0$, and there is no significant difference between the proposed method and the baseline. By contrast, for the alternative hypothesis H_1, $\bar{d} < 0$ and there is a significance difference. Moreover, we report the p_{-value} probability, which helped to determine the significance of the performance of the proposed approach in relation to the null hypothesis. We computed the p_{-value} by evaluating the cumulative distribution function of the t-distribution on the estimated t_{score} value. It is worth noting that the level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that the null hypothesis should be rejected. In our case, the null hypothesis states that the performance of the proposed method and the baselines are similar and they are not statistically significant.

The acceptance or rejection of the null hypothesis depends on a comparison of the p_{-value} with a certain threshold, that is, the critical value. This critical value is the value of the test statistic that defines the upper and lower bounds of a confidence interval. In our case, we set the critical value to 0.05. This means that if the p-value was below this value (e.g., $p_{\text{value}} < 0.05$), we rejected the null hypothesis and concluded that there was a significant difference in performance between the proposed approach and the baselines, and vice versa. In Table 6, we report the computed p_{-value} values. Clearly, we can observe the significant performance of the proposed approach against the baselines.

Table 6
The estimated p_{-value} for interpreting the statistical significance of the proposed method against the baselines. the rejection of the null hypothesis is decided at $p_{-value} < 0.05$. (Yes) means there is a significant difference and (No) means no significant difference.



Method \ Attribute	Be	Ch	Im	Re	Gr
RankNet [29]	0.396 (No)	0.003 (Yes)	0.0404 (Yes)	0.0156 (No)	0.0339 (Yes)
DRA [6]	0 (Yes)	0 (Yes	0 (Yes)	0 (Yes)	0 (Yes)
SQCL [22]	0.0839 (No)	0 (Yes)	0.0005 (Yes)	0.0745 (No)	0.0302 (Yes)
Sinkhorn [7] (No ADUL)	0.0012 (Yes)	0.003 (Yes)	0.0002 (Yes)	0 (Yes)	0.0082 (Yes)
Sinkhorn [7] + ADUL	0.0004 (Yes)	0 (Yes)	0.0026 (Yes)	0 (Yes)	0.0339 (No)



(a) (b)

Fig. 7. Histograms of the ground-truth labels of test pairs (a) and the average accuracies of the corresponding hired annotators (b). The most certain annotators $\{A1, A3, A4, A6, A7, A8\}$ at specific labels $\{-1, 0, 1\}$ achieve best accuracy.

4.7.2. Remarks on performance and time complexity

We observed the high performance of the three-grade scale dataset compared with the seven-grade scale dataset, while the margins of the relative label distribution for the former dataset were not trainable set empirically. In fact, the network parameters, except for the margins of the relative label of the AIUL layer (e.g., parameters in the backbone network that output the mean and standard deviation of the parametric continuous absolute score distribution), were still trainable. Additionally, model performance may have been influenced by other model components, such as the backbone networks, newly added fully connected layers, and ADUL layer (i.e., transition matrix optimization). Furthermore, the three-grade scale dataset was not initially prepared to be used for a distribution-based estimation task. We considered only three labels, which limited the degree of uncertainty and increased the chance of more true positives and high accuracy accordingly. From an algorithmic viewpoint, we argue that estimating the annotator-(in)dependent uncertainties helps us to understand the overall performance of the gait relative attribute estimation system. For crowdsourcing noisy label aggregation, several annotators may be biased toward a specific label, as shown in Fig. 6. In the first row, we can clearly observe from the ground-truth transition matrices that annotators A1, A3, A4, A6, A7, and A8 are certain by the comparative label 0. It means that those annotators agreed that most of the gait pairs they saw had a similar *beautifulness* attribute. For any input test pair, they attempted to assign the label 0. By contrast, if the input test pairs contained a frequent 0 label, we would expect the aforementioned annotators to achieve better accuracies for annotator-specific label distribution estimation. As shown in Fig 7(a), the counts of the ground-truth labels of the testing pairs were concentrated at the comparative label $\{-1, 0, 1\}$. Therefore, we would expect annotators who were certain about these labels to achieve the best accuracy compared with other annotators, as shown in Fig 7(b). Regarding the time complexity, in Table 7, we report the computation time of the proposed method against the baselines. We observed that the proposed method and the baselines almost had the same execution time. This is because we used the same backbone network and the input images were binary silhouettes, where no overhead was required to process the images.

Table 7
The computational time for each input sample (in seconds).

Method \ Attribute	Be	Ch	Im	Re	Gr
RankNet [29]	0.048	0.047	0.047	0.047	0.047
DRA [6]	0.048	0.047	0.047	0.047	0.047
SQCL [22]	0.047	0.047	0.047	0.047	0.047
Sinkhorn [7](No ADUL)	0.049	0.049	0.049	0.048	0.049
Sinkhorn [7] + ADUL	0.050	0.050	0.050	0.050	0.050
Proposed	0.050	0.050	0.050	0.050	0.050

5. Conclusion

In this paper, we introduced an uncertainty-aware estimation model for gait relative attributes. This model can estimate both the annotator-dependent and annotator-independent label distributions of an attribute in one end-to-end task. These estimated distributions well-expressed the underlying uncertainty of the annotation labels. To achieve this, we proposed a differentiable global uncertainty layer module that first estimated the relative score distribution from the absolute score distributions and then mapped it to the relative label distribution. Furthermore, we proposed an annotator-dependent uncertainty layer to learn the underlying uncertainty of each annotator and predict the annotator-specific label distribution through the linear transformation of the relative label distribution. Quantitative and qualitative experiments on two gait relative attribute datasets demonstrated that the proposed model effectively estimated the relative label distribution, annotators' uncertainties, and annotator-specific label distribution in an end-to-end task. The proposed method achieved performance better than and comparable with existing score-based and distribution-based baselines. However, the proposed method may have suffered from poor performance in the case of a low-quality silhouette because of the utilized backbone network, that is, GaitSet. Additionally, we did not attempt to execute this method for RGB input because we aggregated the proposed dataset annotations using a binary silhouette. For real-time applications, our system may not perform well because it reports a relative attribute distribution, which requires a pair of inputs. Instead, we can use a single stream from the trained model to report the absolute attribute distribution or its expectation for the multi-object tracking task. Finally, poor initialization for both the margins of the relative distribution and the transition matrices of the ADUL layer may lead to poor training and performance degradation accordingly. In future work, we will consider extending the model to work on multi-view gait datasets and include more gait relative attributes to make it robust against various covariates, such as view and carrying status. Additionally, we may apply the proposed method to other modalities, such as the face or iris. Furthermore, this model can be applied in several applications, such as walking improvement recommendation systems, gait attribute-based criminal investigations, relative age estimation, medical image quality assessment to support the decision-making of medical staff, and sports action quality assessment to assess judges' scores.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by JSPS Grants-in-Aid for Scientific Research (A) [Grant number JP18H04115]. We thank Maxine Garcia, Ph.D., from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

References

[1] D. Parikh, K. Grauman, Relative attributes, in: IEEE ICCV, 2011, pp. 503–510.
[2] F. Xiao, Y. Jae Lee, Discovering the spatial extent of relative attributes, in: IEEE ICCV, 2015, pp. 1458–1466.
[3] Z. Zhang, Y. Li, Z. Zhang, Relative attribute learning with deep attentive cross-image representation, in: Asian Conference on Machine Learning, PMLR, 2018, pp. 879–892.
[4] B. Siddiquie, R.S. Feris, L.S. Davis, Image ranking and retrieval based on multi--attribute queries, in: IEEE CVPR, 2011, pp. 801–808.
[5] R.N. Sandeep, Y. Verma, C.V. Jawahar, Relative parts: distinctive parts for learning relative attributes, in: IEEE CVPR, 2014, pp. 3614–3621.
[6] X. Yang, T. Zhang, C. Xu, S. Yan, M.S. Hossain, A. Ghoneim, Deep relative attributes, IEEE Trans. Multimed. 18 (9) (2016) 1832–1842.
[7] Y. Makihara, Y. Hayashi, A. Shehata, D. Muramatsu, Y. Yagi, Estimation of gait relative attribute distributions using a differentiable trade-off model of optimal and uniform transports, in: IEEE IJCB, 2021, pp. 1–8.
[8] A. Parkash, D. Parikh, Attributes for classifier feedback, in: ECCV, Springer, 2012, pp. 354–368.
[9] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161.
[10] A. Shehata, Y. Hayashi, Y. Makihara, D. Muramatsu, Y. Yagi, Does my gait look nice? Human perception-based gait relative attribute estimation using dense trajectory analysis, in: ACPR, Springer, 2019, pp. 90–105.
[11] S.A.A. Ahmed, B. Yanikoglu, Relative attribute classification with deep--rankSVM, in: ICPR, Springer, 2021, pp. 659–671.
[12] C. Wan, L. Wang, V.V. Phoha, A survey on gait recognition, ACM Comput. Surv. 51 (5) (2018) 1–35.
[13] J. Lu, Y.-P. Tan, Ordinary preserving manifold analysis for human age and head pose estimation, IEEE Trans. Hum. Mach. Syst. 43 (2) (2012) 249–258.
[14] Y. Makihara, M. Okumura, H. Iwama, Y. Yagi, Gait-based age estimation using a whole-generation gait database, ICB, 2011.
[15] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, Y. Yagi, How confident are you in your estimate of a human age? Uncertainty-aware gait-based age estimation by label distribution learning, in: IEEE IJCB, 2020, pp. 1–10.
[16] L.T. Kozlowski, J.E. Cutting, Recognizing the sex of a walker from a dynamic point-light display, Percept. Psychophys. 21 (6) (1977) 575–580.
[17] J.-H. Yoo, D. Hwang, M.S. Nixon, Gender classification in human gait using support vector machine, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, 2005, pp. 138–145.
[18] S. Yu, T. Tan, K. Huang, K. Jia, X. Wu, A study on gait-based gender classification, IEEE Trans. Image Process. 18 (8) (2009) 1905–1910.
[19] M.A.R. Ahad, T.T. Ngo, A.D. Antar, M. Ahmed, T. Hossain, D. Muramatsu, Y. Makihara, S. Inoue, Y. Yagi, Wearable sensor-based gait analysis for age and gender estimation, Sensors 20 (8) (2020) 2424.
[20] Y. Zhuang, L. Lin, R. Tong, J. Liu, Y. Iwamot, Y.-W. Chen, G-GCSN: global graph convolution shrinkage network for emotion perception from gait, ACCV, 2020.
[21] D. Zhang, Y. Wang, B. Bhanu, Ethnicity classification based on gait using multi-view fusion, IEEE Workshop on Biometrics, 2010.
[22] Y. Hayashi, A. Shehata, Y. Makihara, D. Muramatsu, Y. Yagi, Deep gait relative attribute using a signed quadratic contrastive loss, ICPR, 2021.
[23] C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, J. Lu, The OU-ISIR gait database comprising the large population dataset with age and performance evaluation of age estimation, IPSJ Trans. Comput. Vis. Appl. 9 (1) (2017) 24.
[24] C. Xu, A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, Y. Yagi, J. Lu, Uncertainty-aware gait-based age estimation and its applications, IEEE Trans. Biom., Behav., Identity Sci. 3 (4) (2021) 479–494.
[25] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, J. Zhou, Uncertainty-aware score distribution learning for action quality assessment, in: IEEE CVPR, 2020, pp. 9839–9848.
[26] W. Zhang, K. Ma, G. Zhai, X. Yang, Uncertainty-aware blind image quality assessment in the laboratory and wild, IEEE Trans. Image Process. 30 (2021) 3474–3486.
[27] R. Tanno, A. Saeedi, S. Sankaranarayanan, D.C. Alexander, N. Silberman, Learning from noisy labels by regularized estimation of annotator confusion, in: IEEE CVPR, 2019, pp. 11244–11253.
[28] F. Rodrigues, F. Pereira, Deep learning from crowds, in: AAAI Conference on Artificial Intelligence, vol. 32, 2018.
[29] Y. Souri, E. Noury, E. Adeli, Deep relative attributes, in: ACCV, Springer, 2016, pp. 118–133.
[30] W. Kusakunniran, Attribute-based learning for gait recognition using spatio-temporal interest points, Image Vis. Comput. 32 (12) (2014) 1117–1126.
[31] C. Yan, B. Zhang, F. Coenen, Multi-attributes gait identification by convolutional neural networks, in: International Congress on Image and Signal Processing, IEEE, 2015, pp. 642–647.
[32] D.A. Reid, M.S. Nixon, S.V. Stevenage, Identifying humans using comparative descriptions(2011).
[33] X. Chen, J. Xu, J. Weng, Multi-gait recognition using hypergraph partition, Mach. Vis. Appl. 28 (1–2) (2017) 117–127.
[34] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, Super-fine attributes with crowd prototyping, IEEE Trans. PAMI 41 (6) (2018) 1486–1500.
[35] X. Geng, Label distribution learning, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1734–1748.
[36] J. Wang, X. Geng, Classification with label distribution learning, in: IJCAI, 2019, pp. 3712–3718.
[37] A. Sakata, N. Takemura, Y. Yagi, Gait-based age estimation using multi-stage convolutional neural network, IPSJ Trans. Comput. Vis. Appl. 11 (1) (2019) 4.
[38] Z. Liu, Z. Chen, J. Bai, S. Li, S. Lian, Facial pose estimation by deep learning from label distributions, IEEE CVPR Workshops, 2019.
[39] L. Xu, J. Chen, Y. Gan, Head pose estimation using improved label distribution learning with fewer annotations, Multimed. Tools Appl. 78 (14) (2019) 19141–19162.
[40] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Trans. Image Process. 26 (6) (2017) 2825–2838.
[41] X. Yang, X. Geng, D. Zhou, Sparsity conditional energy label distribution learning for age estimation, in: IJCAI, 2016, pp. 2259–2265.
[42] C. Xing, X. Geng, H. Xue, Logistic boosting regression for label distribution learning, in: IEEE CVPR, 2016, pp. 4489–4497.
[43] K. Su, X. Geng, Soft facial landmark detection by label distribution learning, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 5008–5015.
[44] V. Mnih, G.E. Hinton, Learning to label aerial images from noisy data, in: ICML, 2012, pp. 567–574.
[45] H. Chao, Y. He, J. Zhang, J. Feng, Gaitset: regarding gait as a set for cross-view gait recognition, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8126–8133.
[46] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition, IPSJ Trans. Comput. Vis. Appl. 10 (1) (2018) 1–14.
[47] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980(2014).
[48] C.F. Van Loan, G. Golub, Matrix computations (Johns Hopkins Studies in Mathematical Sciences) (1996).

Allam Shehata received Ph.D. from Egypt-Japan University for Science and Technology in 2018. He is an Assistant Professor at ERI and Researcher at Osaka University. His research interests: vision, pattern recognition. He published in IJCAI, JSEA, CVIU, ACPR, ICPR, and IJCB. He is a reviewer of WACV, ECCV, CVPR, ICCV.

Yasushi Makihara received the B.S., M.S., and Ph.D. from Osaka University. He is a Professor at Osaka University. His research interests are vision, pattern recognition

and imaging. He achieved many awards. He served in responsible positions at top journals and conferences including ACPR, ICCV, CVPR, ECCV, etc.

Daigo Muramatsu received the B.S., M.E., and Ph.D. from Waseda University. He is a professor of the Department of Computer and Information Science, Seikei University. His research interests are pattern recognition, and biometrics including gait recognition. He is a member of IPSJ and IEICE.

Md Atiqur Rahman Ahad, Ph.D. (SM-IEEE, SM-OPTICA) is an Associate Professor of Artificial Intelligence & Machine Learning (Champion, Research & Innovation), Dept. of Computer Science & Digital Technologies, University of East London. He became a Professor at the University of Dhaka (DU) in 2018 and served as a specially appointed Associate Professor at Osaka University (20182022). He works on computer vision, imaging, IoT, healthcare, etc. He was awarded the UGC Gold Medal (handed by the Honorable President of Bangladesh), JSPS Postdoctoral Fellowship, and 40 awards/scholarships. He published 11 books (more to appear soon), 200 journals/conference papers & book chapters. Ahad was invited as keynote/invited speaker 150 times in different conferences/universities. He is an Editorial Board Member, Scientific Reports, Nature.

Yasushi Yagi (Fellow, IPSJ; Senior Member, IEEE) received the Ph.D. from Osaka University. He is a Professor with Osaka University since 2003, former Director of SANKEN, former Executive Vice President of Osaka University, Editorial Board member (IJCV), Vice-President (AFCV). He served as top positions of many top conferences and journals.