Anonymization of Human Gait in Video Based on Silhouette Deformation and Texture Transfer

Yuki Hirose, Kazuaki Nakamura, Member, IEEE, Naoko Nitta, Member, IEEE, and Noboru Babaguchi, Senior Member, IEEE

Abstract—These days, a lot of videos are uploaded onto web-based video sharing services such as YouTube. These videos can be freely accessed from all over the world. On the other hand, they often contain the appearance of walking private people, which could be identified by silhouette-based gait recognition techniques rapidly developed in recent years. This causes a serious privacy issue. To avoid it, this paper proposes a method for anonymizing the appearance of walking people, namely human gait, in video. In the proposed method, we first crop human regions from all frames in an input video and binarize them to get their silhouettes. Next, we slightly deform the silhouettes from the aspects of static body shape and dynamic walking rhythm so that the person in the input video cannot be correctly identified by gait recognition techniques. After that, the textures of the original human regions are transferred onto the deformed silhouettes. We achieve this by a displacement field-based approach, which is training-free and thus robust to a variety of clothes. Finally, the anonymized human regions with the transferred textures are filled back into the input video.

In the results of our experiments, we successfully degraded the accuracy of CNN-based gait recognition systems from 100% to 1.57% in the lowest case without yielding serious distortion in the appearance of the human regions, which demonstrated the effectiveness of the proposed method.

Index Terms—Human gait anonymization, silhouette deformation, static body shape, dynamic walking rhythm, texture transfer, displacement field.

I. INTRODUCTION

NOWADAYS, with the development of the Internet and the spread of camera-equipped devices, many people use video sharing services (VSS) such as YouTube in their daily lives. In VSS, people upload their own video contents as a publisher as well as freely access other people’s contents as a viewer from all over the world. As a result, a massive number of web videos are stored and shared on the Internet, whose amount is still growing. Importantly, these web videos often contain the appearance of private people.

At the same time, techniques of identifying humans by their walking motion, which are so-called gait recognition, have been actively studied and rapidly developed recently [1], [2], [3], [4]. The current mainstream of gait recognition techniques is silhouette-based ones [2], [3], [4], where a human region cropped from an input video is first binarized and then fed into a recognition system to identify who s/he is. The state-of-the-art methods can achieve a good performance even when the resolution of an input human region is quite low.

If gait recognition techniques are maliciously applied to the web videos containing private people, it can be disclosed who they are. This is a serious privacy issue. Moreover, people’s location and behaviors can also be revealed by gait recognition since web videos often include the information of the shooting time, shooting location, and so on. This means that a human’s gait, i.e., her/his walking motion, has become privacy-sensitive information similar with other biological features such as face, voice, fingerprint, and so on. To solve the privacy issue, it is desirable that human regions in the web videos uploaded in VSS should be anonymized by its provider before being published. From the above background, in this paper, we propose a method for anonymizing human gait in a given video.

There are two possible approaches for human gait anonymization: visual abstraction and replacement. In the former, we anonymize the gait of the private people contained in a web video by pixelizing or blurring the corresponding human regions [5]. However, this approach makes the human regions visually unnatural and degrades the quality of the web video. This would not be preferred by the provider of VSS because users visiting to the VSS as a viewer generally favor high quality videos more than low quality ones that may visually frustrates them. Hence, we employ the latter, namely replacement-based approach for gait anonymization, whose specific procedure is as below (see also Fig. 1).

1. Detect and crop the human regions from each frame in an input video.
2. Binarize the cropped regions to obtain their silhouettes.
3. Slightly deform the silhouettes so that gait recognition systems cannot successfully identify who they are.
4. Transfer the texture of the original human regions onto the deformed silhouettes, by which an anonymized version of the human regions is obtained.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
(5) Fill the anonymized human regions back into the input video.

We can easily realize the steps (1) and (2) by using an existing human detection and/or segmentation method. The step (5) is also realized without difficulty by a technique of image inpainting such as [6]. Therefore, in this paper, we particularly focus on the steps (3) and (4) and propose the methods for realizing them. Hereafter, we refer to these two steps as “gait silhouette deformation” and “human region texture transfer”, respectively.

A human gait silhouette is determined by two factors: body shape (including shape of clothes) and posture. The former is static and does not change in a single video, while the latter is dynamic and periodically changes according to each person’s walking rhythm. Both these two factors (i.e., a body shape feature and a sequence of posture features extracted from an input video) are key clues for gait recognition. Hence, the proposed method anonymizes both of them in the process of gait silhouette deformation. To this end, we add a slight perturbation to the shape and posture features extracted from an input video by convolutional neural network (CNN)-based encoders. The perturbed features are then converted to an anonymous gait silhouette by a CNN-based decoder. On the other hand, in the process of human region texture transfer, we do not employ CNNs which require a huge amount of training data to handle a variety of clothes. Instead, we estimate the displacement field [7] between the original and the deformed silhouettes, which is then applied to the original human regions acquired in the first step of the above five-steps approach.

The contributions of this paper are summarized as follows. First, this is the first work focusing on human gait anonymization from the aspect of not only static features (i.e., body shape) but also dynamic features (i.e., walking rhythm). The effect of anonymizing the static and the dynamic features is separately evaluated in our experiments. Second, we propose a general CNN-based framework for gait silhouette deformation that does not impose any restriction on the network structures of the CNN. Third, the proposed method can handle various types, colors, and designs of clothes in the process of human region texture transfer, which is training-free and requires no training data.

In the remainder of this paper, we first review the related work in Section II, especially focusing on the techniques of visual content abstraction and gait recognition. Next, in Section III, we describe the details of the method for gait silhouette deformation, whose key component is the feature perturbation. After that, in Section IV, we describe the method of human region texture transfer in detail, which is based on displacement field estimation as mentioned above. The performance of the proposed method is then experimentally evaluated in Section V. Finally, we conclude this paper in Section VI.

Note that this paper is an extended version of our previous conference paper [8] where we did not focus on human region texture transfer. The differences between the conference paper and this paper are summarized as follows. First, we propose a method of human region texture transfer as well as that of gait silhouette deformation in this paper. The final output of our current method is a colored video of anonymized human regions, whereas the output of the previous method is just a video of anonymized silhouettes. Details of the method of human region texture transfer is described in Section IV. Second, we further improve the method for perturbing shape and posture features, which is described in Sections III-D and III-E. Third, we experimentally evaluate the performance of the proposed method from a broader point of view, whose results are reported in Section V.

II. RELATED WORK

Since human gait anonymization is an emerging topic even in the field of multimedia, there are only a few studies focusing on this novel task. Rather, methods for face anonymization have been actively studied in the past decade, which is deeply related to gait anonymization. Hence, we first review the previous work on face anonymization in Subsection II-A and then move our focus to gait anonymization in Subsection II-B. After that, in Subsection II-C, we review some existing techniques of gait recognition, which is another topic deeply related to this study.

A. Face Anonymization

Visual contents such as images and video often have privacy-sensitive information, whose typical example is human face. Therefore, methods for anonymizing face regions in a visual content have been widely studied.

In classic methods, simple image processing techniques such as blocking-out, pixelization, blurring, and so on are applied to the face regions to anonymize them [9], [10]. This is called visual abstraction, which works well in terms of human perception, whereas it is not always effective for preventing automatic face recognition systems [11], [12], [13]. Once the category and the strength of the visual abstraction filter used for anonymization is revealed, its impact can be easily canceled [12]. Moreover, visual abstraction techniques often yield unnatural appearance.

To cope with the above drawbacks, another kind of face anonymization techniques has been proposed, whose key concept is replacement. Specifically, face regions in a visual content are anonymized by being replaced with other face images. For each face that should be anonymized, Bitouk et al.
proposed to find its nearest neighbors from a pre-constructed face library and use them for the replacement process [14]. Gross et al. proposed k-Same [15] and its extension named k-Same-Select [16], which first divides a set of face images into several clusters so that each cluster has at least k identities, and then replace each face with its cluster center. This is theoretically based on the concept of k-anonymity [17]. Instead of replacing the whole face, Nakashima et al. proposed a patch-wise replacing approach to anonymize face without missing the facial expression in the original face region [18].

More recently, generative adversarial networks (GAN) [19] are often employed for generating an image of realistic yet anonymous face, which is used for the replacement process [20], [21], [22], [23]. GAN-based replacement approach is advantageous in that a well-trained GAN with a sophisticated network structure can explicitly keep not only the facial expressions but also many other attributes such as the age, gender, race, and so on.

B. Gait Anonymization

Compared to face, human gait has attracted less attention from the research field of multimedia anonymization. However, both the visual abstraction and the replacement approaches can be applied to human body regions to anonymize them. Agrawal et al. proposed a method that applies blurring filters to the whole human region in video to anonymize its gait [5]. Mitsugami et al. proposed to overlay a rod-like symbol over each human region appearing in surveillance video to protect the privacy of people [24]. These methods can be classified into the visual abstraction-based gait anonymization technique.

A drawback of the visual abstraction-based techniques is unnatural appearance, which is not desirable for web videos, as mentioned in Section I. Hence, Tieu et al. proposed a replacement-based method of gait anonymization [25]. Their method employs an autoencoder-like CNN that generates an anonymized gait silhouette from two inputs, a target gait silhouette and a noise silhouette. The CNN basically tries to reconstruct the target silhouette, but it also tries to make the output silhouette partially close to the noise silhouette. Thus, the shape of the output silhouette becomes slightly different from that of the target silhouette, by which gait anonymization for the target silhouette is achieved. In later years, Tieu et al. extended their method by introducing a spatial discriminator and a temporal discriminator in order to generate more natural anonymous silhouettes, based on the concept of GAN [26]. They also extended their method in another way to make it robust to low-quality target silhouettes in which several body parts are missing [27].

The focus of Tieu’s methods is to anonymize human gait only from the aspect of the static feature, i.e., body shape. However, we believe that the dynamic feature such as walking rhythm is also individual-specific and should be anonymized. Therefore, in this paper, we focus on both the static and the dynamic features of human gait.

Here, we point out that gait anonymization is somewhat related to adversarial attacks [28], which aim to fool a machine learning-based pattern recognition system without affecting human perception by adding a small random noise to an input pattern. In face anonymization, anonymized face images should be incorrectly recognized by both humans and automated systems. Unlike this, in gait anonymization, it is enough for anonymized gait videos to be incorrectly recognized by only systems since most humans originally do not have the ability of gait recognition. Therefore, gait anonymization might be achieved with adversarial examples.

There have been several existing studies of adversarial attacks against gait recognition [29], [30]. These are trying to add a small noise into a sequence of gait silhouette images; in other words, these adversarial attack methods aim to fool the gait recognition systems at the silhouette level. Since silhouette images are composed only of two kinds of colors, black and white, the noise generated by adversarial attacks becomes salt-and-pepper-like. This kind of noised silhouette is difficult to naturally colorize. This means that adversarial attacks are actually not suitable for gait anonymization. In this paper, we aim to fool gait recognition systems at the colored video level, which is a large difference between our work and adversarial attacks. For this purpose, directly adding a small noise into the original colored video is not a good solution. This is because such a small noise is mostly canceled by a silhouette extraction process performed as a preprocess of gait recognition.

C. Gait Recognition

Gait Recognition is a relatively novel technique of biometric identification. Although the accuracy of gait recognition is currently not comparable to that of well-studied face recognition, it has a unique advantage; that is, gait recognition can be successfully performed even on low resolution videos captured from far away in which human faces cannot clearly be observed. Because of this advantage, gait recognition has been already utilized for the purpose of crime investigation in some countries [3], [4].

There are two categories of gait recognition methods: model-based and silhouette-based. The former uses 3-dimensional human body models such as a surface model and a skeleton model to estimate the posture of the humans in a given video, and identify them based on a sequence of the estimated postures [1]. The performance of the model-based approach heavily depends on video resolution because it is difficult to accurately estimate the human postures from low resolution video. This loses the above advantage. On the other hand, the latter, i.e., the silhouette-based approach, is suitable to the low resolution video.

Silhouette-based gait recognition methods are further divided into two types. One is the methods directly processing a sequence of gait silhouettes to recognize the person in it. Kale et al. proposed this type of method, which utilizes a hidden Markov model to represent the characteristic of each person’s gait silhouette sequence [31]. This method is disadvantageous in computational efficiency. Therefore, the other type of methods has been more actively studied in recent years, which first aggregates all the frames in a given silhouette
sequence into a single frame and uses it as a feature vector or a feature map for the recognition process.

There have been proposed various aggregation methods for the above purpose. Gait energy image (GEI) [32] is a typical one, which is obtained by averaging the silhouettes over one cycle of gait. Because of the averaging process, GEI tends to emphasize motion-less parts in the human body such as the head and the torso. To focus on moving parts such as the hands and the legs, Bashir et al. extended GEI to gait entropy image (GENI) [33], which is obtained by calculating the Shannon-entropy of luminance intensity pixel-by-pixel. Both GEI and GENI loses the information of walking speed and rhythm. To overcome this drawback, Makihara et al. proposed frequency domain features (FDF) [34], which is obtained by pixel-wise discrete Fourier transform along the time axis. With a similar motivation, Setoguchi et al. proposed signed frame difference energy image (SFDEI) [35], which first calculates inter-frame subtraction images (both positive and negative one) between two consecutive frames and then averages them over one cycle of gait. These aggregative images are fed into a CNN as a feature map in modern gait recognition methods [36].

More recently, cross-view gait recognition has become a hot topic. Even the silhouette sequences of the same person’s gait could greatly differ depending on the viewing angle (e.g., frontal view vs. side view), while it is quite difficult in practice to control the viewing angle flexibly for each walking person. Hence, researchers have tried to develop gait recognition methods that are robust to the variety of viewing angles [37],[38],[39]. These advanced methods are powerful yet require a training dataset that contains various gait silhouettes observed from different viewing angles. Note that even these methods assume linear walking since it is rare for ordinary people to suddenly change their walking direction within a narrow area that can be observed by a single camera. This indicates that the viewing angle can be treated as a part of the static component of gait in our context. Thus, we do not explicitly consider the variety of viewing angles in the remainder.

Since there are various features such as GEI and FDF as above, it is desirable that a gait anonymization method can degrade gait recognition accuracy independent from the kind of the features.

III. GAIT SILHOUETTE DEFORMATION BY PERTURBING SHAPE AND POSTURE FEATURES

In this section, we describe the method for gait silhouette deformation in detail, which is achieved by perturbations of shape and posture features extracted from a sequence of human gait silhouettes.

A. Overview

Let \( S = (S_1, \ldots, S_M) \) be a sequence of gait silhouettes that should be anonymized. \( S_i \) is the silhouette image obtained by binarizing the \( i \)-th frame in an input gait video, where \( M \) is the number of the frames. The goal of this section is to deform \( S \) and obtain a new silhouette sequence \( T = (T_1, \ldots, T_M) \) so that gait recognition techniques cannot correctly identify the person in \( T \). To this end, our method achieves the deformation frame-by-frame; that is, we first deform each frame \( S_i \) to \( T_i \) separately and then concatenate the deformed frames into a single video. Therefore, we focus on the frame-wise deformation process in the remainder of this section.

As mentioned in Section I, a gait silhouette is determined by two factors: shape (the shape of body and that of clothes) and posture. The former represents the static aspect of the input gait while the latter represents the dynamic aspect. Basically, the same person’s silhouettes always have the same shape in a single video regardless of posture, i.e., frame ID. On the other hand, posture is different in individual frames, which can be expressed by “phase” \( \theta \in [0, 2\pi) \) because of the periodicity of walking motion. Based on the above discussion, we can re-write the frame \( S_i \) as \( S_i = \text{Sil}_a(\theta_i) \), whose meaning is the gait silhouette of a person \( a \) with phase \( \theta_i \). We can anonymize the \( \text{Sil}_a(\theta_i) \) from both the static and the dynamic aspects by changing it to \( \text{Sil}_a'(\theta'_i) \), where \( \theta' \) is a fictional person different from \( a \) and \( \theta'_i \) is a different value from \( \theta_i \).

The concrete procedure of the proposed deformation method is as follows.

(i) A “shape code” \( z_a = E[\text{Sil}_a(\theta_i)] \) which is a feature vector representing the shape of \( \text{Sil}_a(\theta_i) \), is extracted by a certain encoder \( E \). At the same time, the value of the phase \( \theta_i \) is estimated in some way from the \( \text{Sil}_a(\theta_i) \).

(ii) Perturbations \( \Delta z \) and \( \Delta \theta \) are respectively added to the shape code and the phase value to get \( z_a' = z_a + \Delta z \) and \( \theta'_i = \theta_i + \Delta \theta \).

(iii) A new silhouette image \( \text{Sil}_a'(\theta'_i) = D[z_a', \theta'_i] \) is generated from the perturbed shape code \( z_a' \) and phase value \( \theta'_i \) by a certain decoder \( D \).

The silhouette image obtained in the third step is finally used as the anonymized version of the frame \( S_i \), that is, \( \text{Sil}_a'(\theta'_i) = T_i \). Note that the perturbation for the shape code is always same for all \( i \in \{1, \ldots, M\} \) to avoid unnatural shape change in \( T \).

Hereafter, we describe how to estimate the phase value in Subsection III-B and how to train the encoder \( E \) and the decoder \( D \) in Subsection III-C. Then, the strategies for determining the perturbations \( \Delta z \) and \( \Delta \theta \) are described in detail in Subsections III-D and III-E, respectively. For convenience of explanation, we assume that all gait silhouette sequences used in the subsequent subsections only include just one cycle of walking motion, which can be easily extracted by using an autocorrelation analysis as a pre-process. In practice, there are cases where the length of an input sequence is longer than one cycle, of course. In these cases, we extract the first cycle from the input sequence and apply the proposed method to it. Then, for every remaining frame with a phase value \( \theta_i \geq 2\pi \), we anonymize it (i.e., the \( i \)-th frame) in the same way as the anonymization process of the \( j \)-th frame which is the nearest neighbor of the \( i \)-th frame in terms of phase; that is, the perturbations \( \Delta z \) and \( \Delta \theta \) that are used to anonymize the \( j \)-th frame are also used for the \( i \)-th frame, where \( 0 \leq \theta_j \approx \theta_i - 2\pi n < 2\pi \) with a certain integer \( n \). For a gait sequence whose length is shorter than one cycle, we do not have to anonymize it. This is because the gait recognition
systems require more than or equal to one cycle of gait to accurately recognize it.

B. Defining and Estimating Phase Value

Before explaining how to estimate the phase value, we first have to define it clearly.

Naively, the phase values of the first and the last frame in one cycle of gait video are defined as 0 and 2π, respectively, and the phase values of the intermediate frames are defined by linear interpolation, as shown in Fig. 2 (a). However, this is not a desirable definition because different values could be given to the frames having the same posture when there are two or more videos. It is an important requirement to give the same phase value to all the frames having the same posture, regardless of individual people, as shown in Fig. 2 (b). On the other hand, any definition is allowed as long as this requirement is satisfied. For example, the definition shown in Fig. 2 (c) is also a desirable one. Based on the above consideration, we prepare a certain reference sequence \( R = (R_1, \ldots, R_N) \) to define the phase value, where \( R_i \) is the \( i \)-th frame in \( R \) and \( N \) is its length. For each \( R_i \), we define its phase value as \( 2\pi \frac{i-1}{N} \).

According to the above definition, we estimate the phase values of any other gait silhouette sequence \( S = (S_1, \ldots, S_M) \) by making a correspondence between \( R \) and \( S \) via DP matching. However, if we directly apply DP matching to \( R \) and \( S \), \( S_j \) always matches \( R_1 \), although the posture of the person in \( S_1 \) is not always same with that in \( R_1 \). Hence, we first make a periodic shifted version of \( S \) as

\[
S_l = (S_{l+1}, S_{l+2}, \ldots, S_M, S_1, S_2, \ldots, S_l),
\]

and then we apply DP matching to \( R \) and \( S_l \). Let \( C(R, S_l) \) be the matching cost. We perform the above process for all \( l \in \{0, 1, \ldots, M - 1\} \) and find the best \( \hat{l} \) that minimizes the cost \( C(R, S_{\hat{l}}) \), that is,

\[
\hat{l} = \text{argmin}_l C(R, S_l).
\]

Based on the matching result between \( R \) and \( S_{\hat{l}} \), we estimate the phase value of \( S_j \) as \( 2\pi \frac{(j-1)}{M} \) if \( S_j \) was matched to \( R_1 \). See Appendix A for more details.

C. Training Gait Silhouette Encoder and Decoder

As mentioned in Subsection III-A, we use an encoder \( E \) to extract a shape code \( z_a \) from \( \text{Sil}_a(\theta_i) \) as well as use a decoder \( D \) to generate a gait silhouette image \( \text{Sil}_s(\theta'_i) \). The proposed framework for achieving these processes is briefly shown in Fig. 3.

D. Shape Perturbation by Quadratic Optimization

For any phase value \( \theta_i \), we convert it to a two-dimensional vector \( \mathbf{p}_\theta = (\cos \theta \sin \theta) \) before feeding it into the decoder \( D \). This allows us to equally treat the phase values 0 and 2π despite their gap on the real number line.

In the training process for \( E \) and \( D \), we do not add any perturbation to the shape code and the phase value. In this case, \( E \) and \( D \) behave like an autoencoder and therefore we can employ the mean squared error (MSE) between an input and an output silhouette images as a loss function. However, \( E \) should output the same shape code \( z_a \) from \( \text{Sil}_a(\theta_i) \) for all \( \theta_i \) unlike autoencoders. To ensure this property under the strategy of using the MSE loss, we have to define the ground-truth of \( z_a \) in some way. To this end, we employ the following approach. First, using a lot of gait silhouette sequences as a training dataset, we train a variational autoencoder (VAE). Let \( E_{\text{vae}} \) be the encoder part of the trained VAE. Next, each silhouette image \( \text{Sil}_a(\theta_i) \) in the training dataset is compressed to a feature vector \( \xi_a(\theta_i) = E_{\text{vae}}[\text{Sil}_a(\theta_i)] \), whose average with respect to \( \theta_i \) is calculated and used as the ground truth of \( z_a \), i.e.,

\[
z_{a}^{\text{gt}} = \frac{1}{M_a} \sum_{i=1}^{M_a} \xi_a(\theta_i).
\]  

\( M_a \) is the length of the gait silhouette sequence of the person \( a \) in the training dataset.

Using the above \( z_{a}^{\text{gt}} \), we simultaneously train \( E \) and \( D \) by minimizing

\[
L(E, D) = \sum_a \sum_{i=1}^{M_a} \left[ \| E[\text{Sil}_a(\theta_i)] - z_{a}^{\text{gt}} \|^2 + \lambda \| D[E[\text{Sil}_a(\theta_i)], \mathbf{p}_\theta] - \text{Sil}_a(\theta_i) \|^2 \right],
\]

where \( \lambda \) is a weighting constant to control the balance between the first and the second terms. Note that the dataset for training \( E \) and \( D \) is totally same with the one used for training the VAE. See also Appendix B for more details on the training process of the \( E \) and \( D \).
under the constraint (7), where the second term corresponds to determining as \( z \) whose elements are 1, and the other is \( I \) because of too large appearance would be caused in the output gait silhouette a problem as mentioned above; namely, seriously distorted that is definitely not close to \( z \). To prevent this, we give the anonymization capability, we naively try to find an equivalent with determining as \( c \). Similarly, let \( e \) be the vector whose \( k \)-th element is \( c_k \), i.e., \( e = (c_1 \cdots c_K)^T \). Using \( Z \) and \( e \), the perturbed shape code \( z_{a'} \) in Formula (5) can be re-written as \( z_{a'} = Z e \). This is equivalent with determining as \( \Delta z = Z e - z_a \).

To make \( ||\Delta z|| \) as large as possible to get enough anonymization capability, we naively try to find

\[
\hat{c} = \arg\max_c ||\Delta z||^2 = \arg\max_c ||Ze - z_a||^2, \tag{6}
\]

and calculate \( z_{a'} \) as \( z_{a'} = Ze \hat{c} \). This allows us to obtain \( z_{a'} \) that is definitely not close to \( z_a \). However, this solution has a problem as mentioned above; namely, seriously distorted appearance would be caused in the output gait silhouette because of too large \( ||\Delta z|| \). To prevent this, we give the following two constraints on the \( c \). One is

\[
\sum_{k=1}^{K} c_k = 1 \iff e^T \mathbb{1} = 1, \tag{7}
\]

where \( \mathbb{1} = (1 \cdots 1)^T \) is the \( K \)-dimensional vector all of whose elements are 1, and the other is

\[
\forall k \in \{1, \cdots, K\} \quad 0 \leq c_k \leq 1. \tag{8}
\]

These constraints are expected to keep \( z_{a'} \) close to at least one of \( \{z_{a,k}\}_{1 \leq k \leq K} \), achieving natural appearance in the output gait silhouette. For convenience of computation, we do not directly consider the second constraint (8). Instead, we introduce it as a regularization term into the objective function (6); that is, we actually find

\[
\hat{c} = \arg\max_c \left\{ \eta ||Ze - z_a||^2 + \sum_{i=k}^{K} c_k(1 - c_k) \right\}
\]

\[
\arg\max_c \left\{ \eta ||Ze - z_a||^2 + e^T (\mathbb{1} - e) \right\} \tag{9}
\]

under the constraint (7), where the second term corresponds to the constraint (8) and \( \eta \) is a weighting constant.

Since the above optimization problem (9) is quadratic with respect to \( e \), its solution is easily obtained as

\[
\hat{e} = q + \frac{1}{2} \pi^T q h \tag{10}
\]

by the method of Lagrange multiplier, where \( q = \eta G^{-1} Z^T z_a \), \( h = G^{-1} \mathbb{1} \), and \( G = Z^T Z - I_K \). \( I_K \) is the \( K \times K \) identity matrix. Note that the quadratic term of the objective function (9) is

\[
\eta e^T Z^T Z e - c^T e = c^T G e, \tag{11}
\]

which means the above maximization problem can be correctly solved if and only if \( G \) is negative semi-definite. To ensure this property, \( \eta \) should be smaller than \( \frac{1}{\tau} \), where \( \tau \) is the largest eigenvalue of \( Z^T Z \). Hence, we set \( \eta = \frac{\nu}{\tau} \) in our experiments by introducing a parameter \( \omega \) that satisfies 0 \( \leq \omega < 1 \). The setting of \( \omega = 1 \) makes \( G \) singular, in which \( G^{-1} \) cannot be computed.

E. Phase Perturbation Based on Divergence Maximization

The dynamic aspect of human gait is represented by not a single phase value but its sequence. Hence, the phase perturbation \( \Delta \theta_i \) should not be separately determined frame-by-frame. Instead, our proposed method simultaneously determines the \( \Delta \theta_i \) for all \( i \) \( (1 \leq i \leq M) \).

In general, a person’s walking direction is consistent and her/his posture changes continuously in a single video. Hence, a sequence of phase values \( (\theta_1, \cdots, \theta_M) \), which represents one cycle of walking motion, satisfies the following equation:

\[
\sum_{i=1}^{M} \phi_i = 1 \quad \text{where} \quad \phi_i = \frac{1}{2\pi} (\theta_i - \theta_{i-1}) \mod 2\pi. \tag{12}
\]

Note that we introduce \( \theta_0 = \theta_M \) for sake of convenience. Since \( \theta_i \geq 0 \) is also satisfied for all \( i \), we can consider that \( \Phi = (\phi_1, \cdots, \phi_M) \) is virtually a probability distribution (see Fig. 4). Similarly, for the phase value sequence after perturbation \( (\theta'_1, \cdots, \theta'_M) \), we can consider that \( \Phi' = (\phi'_1, \cdots, \phi'_M) \) is a probability distribution, where

\[
\phi_i' = \frac{1}{2\pi} (\theta_i' - \theta_{i-1}') \mod 2\pi. \tag{13}
\]
To anonymize the dynamic feature of the input sequence, we should make $\Phi'$ as dissimilar with $\Phi$ as possible.

To measure the dissimilarity between two probability distributions $\Phi$ and $\Phi'$, we employ Jensen-Shannon (JS) divergence, which is calculated as

$$\text{JS}(\Phi||\Phi') = \frac{1}{2} \sum_{i=1}^{M} \left( \phi_i \log \frac{2\phi_i}{\phi_i + \phi'_i} + \phi'_i \log \frac{2\phi'_i}{\phi_i + \phi'_i} \right).$$

(14)

The optimal $\Phi'$ that maximizes $\text{JS}(\Phi||\Phi')$ can be obtained as

$$\phi'_i = \left( \frac{(\phi_i)^\alpha}{\sum_{j=1}^{M} (\phi_j)^\alpha} \right)$$

for all $i$ in the limit as $\alpha \to -\infty$. However, this solution also maximizes the bias of the distribution $\Phi'$, which yields discontinuous posture change in the output gait silhouette sequence. This problem is avoidable by setting $\alpha = 0$, which results in $\phi'_i = \frac{1}{M}$ for all $i$, but the anonymization capability is degraded. Thus, we regard the $\alpha$ as a hyperparameter and empirically search its optimal value in our experiments mentioned in Section V-B.

Using the optimal $\Phi'$ found by the above process, we compute the perturbed phase values as

$$\theta'_i = \theta_0 + 2\pi \sum_{j=1}^{i} \phi'_j,$$

(15)

which is equivalent with determining as $\Delta \theta_i = 2\pi \sum_{j=1}^{i} (\phi'_j - \phi_j)$.

**IV. HUMAN REGION TEXTURE TRANSFER BASED ON DISPLACEMENT FIELD ESTIMATION**

In this section, we explain how to transfer the texture of original human regions onto the corresponding deformed silhouettes obtained in the previous section.

**A. Overview**

In the modern methods of generating photo-realistic human action video [40], [41], a CNN-based encoder-decoder architecture is often used for adjusting the texture of a human region in a source image/video to arbitrary target poses. This architecture is theoretically applicable to the task of the texture transfer on which we focus. However, it cannot handle a variety of clothes without a huge amount of training data that can be hardly collected. Hence, we employ a training-free method for the texture transfer based on the idea of displacement fields.

A displacement field (DF) is a two-dimensional vector field representing pixel-wise displacement vectors between two fixed-size images $J(x, y)$ and $I(x, y)$, where $J$ and $I$ are only slightly different from each other. When a pixel $(x, y)$ on the first image $J$ is corresponding to the pixel $(u, v)$ on the other image $I$, their displacement vector is defined as $(u-x, v-y)$. Since both the horizontal component $u-x$ and the vertical component $v-y$ vary depending on the pixel location, we express them as $f(x, y)$ and $g(x, y)$, respectively. This means that the images $J$ and $I$ satisfy

$$J(x, y) \approx I(u, v) = I(x + f(x, y), y + g(x, y)),$$

(17)

for all $(x, y)$, where a pair of $f$ and $g$ is a DF. A typical example of DFs is an optical flow field densely computed between two consecutive frames in video.

With the method proposed in Section III, we have already obtained a sequence of deformed gait silhouettes $T$ from an input sequence $S$. Let $T_i(u, v)$ be the $i$-th frame in $T$ and let $S_i(u, v)$ be the $i$-th frame in $S$. In addition, let $I_i(u, v)$ be the original colored version of $S_i(u, v)$. In this situation, once an appropriate DF between $T_i$ and $S_i$ is obtained, we can easily transfer the texture of $I_i$ onto $T_i$ and get its colored version $J_i$ by Algorithm 1, where $f$ and $g$ are the DF satisfying $T_i(x, y) \approx S_i(x + f(x, y), y + g(x, y))$, as shown in Fig. 5. $H$ and $W$ are the image height and width, respectively. Thus, how we estimate the appropriate $f$ and $g$ from a pair of $T_i$ and $S_i$ plays a key role here, which is more difficult if the phase value of $S_i$ is more different from that of $T_i$. Hence, it is also important to select in-phase pairs of silhouette images as $T_i$ and $S_i$. We describe how to achieve these two sub-tasks in the subsequent subsections.

**B. Selection of In-Phase Silhouette Image Pairs**

For each $T_i \in T = \{T_1, \cdots, T_M\}$, we have to select the best counterpart for it from $S = \{S_1, \cdots, S_M\}$. This is achieved by calculating the silhouette similarity between $T_i$ and $S_i$ for all $i \in \{1, \cdots, M\}$ and then selecting the most similar one. Separately performing this process for each $i \in \{1, \cdots, M\}$ is the simplest way. However, this strategy often causes a perceptible flicker in the output image sequence $\{J_1, \cdots, J_M\}$. This is because the counterpart of
$T_i$ and that of $T_{i+1}$ are not always temporally neighbored with each other in $S$. To avoid the flicker effect, we employ the silhouette matching method same with the one mentioned in Section III-B. More specifically, we find the best correspondence between $T$ and $S$ by using DP matching and the periodic shift operator. Based on its result, we select the counterpart of $T_i$ for each $i$.

C. Estimation of Appropriate Displacement Field

For each pair of in-phase silhouette images $T_i$ and $S_i$ selected in the previous subsection, we estimate the DF between them. Basically, this can be achieved by finding the $f$ and $g$ that minimizes the cost function

$$Q_1(f, g) = \sum_{(x,y) \in A} [T_i(x, y) - S_i(x+f(x, y), y+g(x, y))]^2,$$

(18)

where $A$ is the whole region in the image $T_i$. However, since both the $T_i$ and $S_i$ are a binary image whose pixel values are either 0 or 1, the above cost $Q_1$ easily becomes 0 with a discontinuous, unnatural, and large-magnitude DF. To avoid this, we introduce the following three constraints.

**Smoothness constraint** is first introduced for guaranteeing the spatial smoothness of $f$ and $g$. If these two functions are spatially smooth, their values do not drastically change between two adjacent pixels, resulting in a continuous DF. This is generally achieved by minimizing the second-order derivative of $f$ and that of $g$, which are given as

$$L_f(x, y) = f(x-1, y) + f(x+1, y) + f(x, y-1) + f(x, y+1) - 4f(x, y)$$

(19)

and $L_g(x, y)$, respectively, with the Laplacian operator $L$. Hence, we minimize

$$Q_2(f, g) = \sum_{(x,y) \in A} [(L_f(x, y))^2 + (L_g(x, y))^2]$$

(20)

in addition to $Q_1$.

**Consistency constraint** is introduced for guaranteeing the consistency of pixel alignments between $T_i$ and $S_i$. Suppose that the pixels $(x, y)$ and $(x+1, y)$ on $T_i$ are corresponding to the pixels $(u, v)$ and $(u', v')$ on $S_i$, respectively. If $u' = f((x+1), y) + f(x, y+1)$ is larger than $u = x + f(x, y)$, in other words, if $u - u' = -1 + f(x, y) - f(x+1, y) < 0$ is satisfied, the order of these two pixels is consistent between the two images. This is a key property for making $f$ natural, which is also the case with $g$. Hence, we minimize

$$Q_3(f, g) = \sum_{(x,y) \in A} [(C_f(x, y))^2 + (C_g(x, y))^2]$$

(21)

as well as $Q_1$ and $Q_2$, where

$$C_f(x, y) = \max(0, -1 + f(x, y) - f(x+1, y))$$

(22)

and

$$C_g(x, y) = \max(0, -1 + g(x, y) - g(x, y+1)).$$

(23)

V. EXPERIMENTS

We evaluated the effectiveness of the proposed method with several experiments, whose results are reported in this section. We first summarize the setup of the experiments in Subsection V-A and describe how to tune the hyperparameters in Subsection V-B. Then, we show the experimental results and give some discussions on them in Subsection V-C.

A. Experimental Setup

1) Dataset: In our experiments, we used two types of gait videos. One is gait silhouette videos and the other is colored gait videos.

For the former, we employed the treadmill dataset A and B from the OU-ISIR Gait Database [42]. The treadmill dataset A includes 612 gait silhouette videos of 34 people, whose walking speed is ranged from 2 [km/h] to 10 [km/h]. However, since some of them look unordinary, i.e., too slow or too fast, we used only 204 videos in which a person is walking at the speed of 4, 5, or 6 [km/h]. We refer to a set of these 204 videos as DS$_a$ in the remainder. Meanwhile, the treadmill dataset B includes 2176 gait silhouette videos of 68 people whose walking speed is ordinary enough. Thus, we used all of them, which we call DS$_b$ in the remainder.

For the latter, namely colored gait videos, we constructed our own dataset by shooting 14 people with a web camera 13 or 14 times for each person. The number of the collected videos is 190, where the human region in each frame was

![Fig. 6. Whole image region A and its boundary region B.](image-url)
extracted with a chroma-key system and the remaining background region was filled with gray color. This is a pre-process for easily binarizing the colored videos to obtain their silhouettes. We refer to this dataset as DS\textsubscript{a} in the remainder. The resolution of the videos in DS\textsubscript{a} was adjusted to 88 x 128 pixels, which is identical to the resolution of the videos in DS\textsubscript{b} and DS\textsubscript{c}.

Note that all the videos in the above datasets include just one cycle of linear walking as mentioned in Subsection III-A.

2) Training and Testing Procedure: First, in the training phase, we used DS\textsubscript{b} as a training dataset to train the encoder \(E\) and the decoder \(D\) mentioned in Subsection III-C. In addition, we also used DS\textsubscript{b} with the trained encoder \(E\) to construct the shape code database, which is needed to determine the shape perturbation in the way of Subsection III-D. After that, in the testing phase, we tried to anonymize 64 gait videos in DS\textsubscript{c} with the proposed method, where the remaining 126 videos were used for another purpose as described later.

To separately evaluate the effect of the shape perturbation and that of the phase perturbation, we actually tested the following three methods and compared them: \textit{shape-only}, \textit{phase-only}, and \textit{both}. In the \textit{shape-only} method, we did not add any phase perturbation into the input gait videos. Similarly, in the \textit{phase-only} method, we did not add any shape perturbation. In the \textit{both} method, we added shape perturbation as well as phase perturbation.

3) Evaluation Criteria: The anonymization results were evaluated from two aspects: anonymization performance and visual naturalness.

For anonymization performance, we adopted gait recognition accuracy as an evaluation criterion. Lower accuracy means higher anonymization performance when we input anonymized gait videos into a certain gait recognition system. For this purpose, we constructed GEINet-based gait recognition systems. Although there are more advanced gait recognition methods that can handle the variety of viewing angles as mentioned in Subsection II-C, these are not suitable for the OU-ISIR Gait Database, which is a database consisting only of side-view gait silhouettes. This is also the case with our own dataset DS\textsubscript{c}. Hence we did not use such cross-view methods. GEINet \cite{36} is a neural network for gait recognition, which is composed of two sequential triplets of convolution, pooling, and normalization layers and two subsequent fully-connected layers. Originally, GEINet is assumed to take GEI of a gait silhouette video as an input feature map. However, GEI loses dynamic information of the walking action, as mentioned in Subsection II-C. Therefore, we separately constructed three networks each of which takes GEI, FDF, and SFDEI as an input feature, respectively. The structure of these three networks is same as that of the original GEINet except for the number of channels in the input layer. The concrete process to train the networks is as follows. First, we binarized the 126 videos in DS\textsubscript{a} that are not used in the testing phase, whose resultant silhouettes were used as training samples. In addition, to cover more people as target classes, we also used all the videos in DS\textsubscript{a} as training samples. Since DS\textsubscript{c} contains videos of 14 people while DS\textsubscript{a} contains videos of 34 people, the trained networks cover 48 classes (or 48 people) in total.

After training the three GEINet-based networks, we anonymized the remaining 64 videos in DS\textsubscript{a} as mentioned above. Then we evaluated the anonymization results in the following two ways. One is to directly input the results of gait silhouette deformation proposed in Section III to the three networks before performing the human region texture transfer (HRTT) proposed in Section IV. The other is to input the re-binarized version of the final anonymization results after performing HRTT. We refer to these two ways as \textit{before-HRTT} and \textit{after-HRTT} in the remainder of this section.

For visual naturalness, it is not easy to objectively evaluate the gait videos from such an aspect. Tieu \textit{et al.} employed a human action recognition system for a similar purpose \cite{26}. This is based on the consideration that gait videos having visually natural motion tend to be correctly recognized as “walking” by action recognition systems whereas videos with unnatural motion tend to be wrongly recognized. Inspired by them, we employed a pre-trained CNN model of action recognition called 3D-ResNet \cite{43}. In addition, we also employed an object detection model called YOLO \cite{44} under the consideration that human regions in the gait videos having a visually natural appearance tend to be correctly detected by YOLO. We input the gait videos anonymized by the proposed method into not only 3D-ResNet video-by-video but also YOLO frame-by-frame and measured their recognition/detection accuracy, where higher accuracy indicates better visual naturalness.

B. Hyperparameter Tuning

As introduced in Subsections III-D and III-E, we have several hyperparameters which could significantly affect the anonymization performance of the proposed method as well as the visual naturalness of its resultant gait videos. They are, namely, \(\omega\) and \(K\) for the shape perturbation and \(\alpha\) for the phase perturbation. We first examined the impact of these hyperparameters to tune them, whose results are reported and discussed in this subsection. As described later in Subsection V-C, we finally employed \(\omega = 0.99\), \(K = 20\), and \(\alpha = -0.5\) through the examination. However, we only reported the cases when two of them were fixed and the remaining one was varied, due to page limitations.

The hyperparameters for human region texture transfer, namely \(\beta_2\), \(\beta_3\), and \(\beta_4\) in Formula (25) were empirically set as \(\beta_2 = 0.3\), \(\beta_3 = 10\), and \(\beta_4 = 5\).

1) Hyperparameters for the Shape Perturbation: In the proposed method, shape perturbation is achieved by replacing the original shape code with a linear combination of its \(K\)-nearest neighbors, whose coefficients are determined by solving the optimization problem (9). \(\omega\) is a weighting constant for balancing the first and the second terms in Formula (9). Their impacts were first examined, whose results are shown in Fig. 7 and Fig. 8. The horizontal axis in these figures indicates GEI-based gait recognition accuracy when \textit{before-HRTT} silhouettes obtained by the \textit{shape-only} strategy were used, and the vertical axis indicates how accurately the anonymized gait videos are recognized as the “walking” action by 3D-ResNet. This means the left area in the figure is desirable.
Fig. 7. Impact of $K$ on anonymization performance and visual naturalness of resultant gait video.

Fig. 8. Impact of $\omega$ on anonymization performance and visual naturalness of resultant gait video.

Fig. 9. Impact of $\alpha$ on anonymization performance and visual naturalness of resultant gait video.

from the aspect of anonymization performance while the upper area is desirable from the aspect of visual naturalness; each hyperparameter is desirable to be tuned so that its result is as close to the top-left corner as possible.

In Fig. 7, we can see that higher $K$ basically leads to higher anonymization performance, especially in the cases with $K \leq 20$. On the other hand, from the aspect of visual naturalness, $K$ does not make a significant effect; the accuracy of 3D-ResNet only slightly changes with $K$ compared to the gait recognition accuracy. Here, we employed $K = 20$, which seems to be closest to the top-left corner of Fig. 7.

The same tendency is also observed in Fig. 8; that is, larger $\omega$ leads to higher anonymization performance, while it does not make a significant impact on the visual naturalness. $\omega = 0.99$ or 0.875 is best since they are clearly closest to the top-left corner of Fig. 8. Thus, we employed $\omega = 0.99$.

2) Hyperparameter for the Phase Perturbation: In the phase perturbation, the hyperparameter $\alpha$ in Formula (15) controls the difference between the original phase sequence and the perturbed one. Theoretically, the difference is maximized by $\alpha \rightarrow -\infty$, whereas it would also degrade the visual naturalness of the anonymized gait videos. We experimentally examined this, whose result is shown in Fig. 9. The meanings of the horizontal and vertical axes are almost the same as those in Fig. 7 and Fig. 8, except for the use of the phase-only strategy instead of the shape-only. As we theoretically considered, $\alpha$ with a larger absolute value leads to higher anonymization performance as well as lower visual naturalness. Especially, in the cases when $\alpha < -0.5$, the visual naturalness is rapidly degraded. This is because such $\alpha$ makes it difficult to successfully achieve human region texture transfer. For these reasons, we employed $\alpha = -0.5$ to strike a balance between the anonymization performance and the visually natural appearance.

C. Results and Discussion

Using the hyperparameters tuned as above, we anonymized the test videos in DS, and evaluated them qualitatively and quantitatively.

1) Qualitative Evaluation: Fig. 10 shows two examples of original gait videos (non-anonymized ones) and their anonymized versions obtained by the proposed method. We can see from this figure that the anonymized videos can keep as natural appearance as their original. There are no serious differences between the original and the anonymized videos, even in the case of the both strategy. The same property can also be observed in most of the other test data, by which it is demonstrated that our proposed method does not degrade the visual naturalness of input gait videos.

Of course, the visual naturalness of the anonymization result is affected by the performance of HRTT, which depends on the three constraints introduced in Subsection IV-C. We qualitatively verified their effect. Fig. 11 shows an example of the displacement field (DF) estimation results and the final anonymization results with and without each constraint. We can see from Fig. 11 that a discontinuous DF is obtained if we drop the smoothness constraint, which causes unnatural texture distortion on the final anonymization result. Especially,
the head region is seriously distorted, which indicates the importance of this constraint. On the other hand, the lack of consistency constraint does not cause any serious distortion. This is because the consistency constraint can be implicitly satisfied by the use of smoothness constraint if the magnitude of the estimated DF is low. Since we use in-phase silhouette pairs for the DF estimation, a low-magnitude DF is obtained in most
cases. We can conclude that the consistency constraint is not necessarily needed. At last, without the boundary constraint, relatively large displacement vectors are obtained around the boundary region, whose texture could be unnaturally stretched. Hence, the boundary constraint is also important, as is the smoothness constraint. Fig. 12 shows the convergence trend of each term in Formula (25), where $Q_3$ is much smaller than the other terms. This also indicates the importance of the smoothness constraint and the boundary constraint compared to the consistency constraint.

2) Quantitative Evaluation in Terms of Anonymization Performance: TABLE I shows the gait recognition accuracy when the original and the anonymized gait videos were input into the GEINet-based gait recognition systems. For comparison, we also tested two visual abstraction-based approaches, namely pixelization and blurring, in addition to the proposed method. Moreover, Fig. 13 shows each person’s accuracy individually, under the conditions of the use of the GEI feature and before-HRTT.

Without any anonymization processes, all the gait features (GEI, FDF, and SFDEI) can achieve a recognition accuracy of 100%. This indicates the necessity of an anonymization process. In contrast, with our proposed method, the gait recognition accuracy is greatly reduced, which shows its high anonymization capability. Compared to the cases of shape-only and both, the anonymization performance of phase-only is limited. This indicates that the dominant factor for recognizing human gait is its static features rather than dynamic features. This is why even GEI, which loses the dynamic features, can achieve the accuracy of 100% in the case of non-anonymized. However, phase perturbation is still useful because the both strategy allows us to reduce the gait recognition accuracy more than the shape-only strategy for almost all the people with ID:1 to ID:14, especially in the cases of FDF and SFDEI. These two features can handle the dynamic aspect of human gait, unlike GEI. Since the phase perturbation can anonymize the dynamic aspect, it is more effective for deceiving the dynamic feature-based gait recognition systems such as FDF and SFDEI.

When comparing the results of before-HRTT and after-HRTT silhouettes, the latter achieves less anonymization performance. This indicates that HRTT sometimes degrades the anonymization capability. This might be due to the smoothness constraint introduced in Section IV-C. The smoothness constraint weakens the high-frequency components in the estimated DF, by which the local shape information of the human regions in the original video might be reflected in the anonymized one, as shown in Fig. 14. Even though the effect of the smoothness constraint is not so serious, it is not desirable, because attackers who want to reveal some privacy information from web videos can input a re-binarized version of anonymized human gait video into their gait recognition systems. In our future work, we will reconsider how to appropriately impose the smoothness constraint on the DF estimation process.

The visual abstraction-based approaches, namely pixelization and blurring, cannot reduce the gait recognition accuracy so much compared to our method. Especially, an accuracy of around 80% is still kept in the case of pixelization. This is because CNN-based gait recognition systems such as GEINet have an internal process similar to pixelization (e.g. average pooling) and therefore are robust to pixelized images.

There is a concern that data augmentation by slight silhouette deformation might enhance the capability of the gait recognition system so that it can defeat our anonymization method. To address the concern, we also tested the effect of data augmentation on gait recognition accuracy. Specifically, we applied dilation, erosion, and scaling operators to the human silhouette in each training sample in DS$_a$ and DS$_c$, and then re-trained the GEINet-based gait recognition systems. For the dilation and erosion operators, we used the 3 × 3 rectangle kernel. For the scaling operator, we used the following four ways to set the horizontal scaling factor $\tau_x$ and the vertical one $\tau_y$:

- $\tau_x = 1.0, \tau_y = 0.95$,
- $\tau_x = 1.0, \tau_y = 1.05$,
- $\tau_x = 0.95, \tau_y = 1.0$,
- $\tau_x = 1.05, \tau_y = 1.0$.

Table II shows the result. The anonymization performance of shape-only is slightly decreased, however, that of phase-only and both is not affected by the data augmentation. This demonstrates that the phase perturbation, which is the main advantage of our proposed method over existing ones, is robust to the data augmentation-based gait recognition.

3) Quantitative Evaluation in Terms of Visual Naturalness: As mentioned in Subsection V-A, the visual naturalness of the anonymized gait videos was evaluated by two criteria: frame-by-frame person detection accuracy by YOLO and video-by-video “walking” action recognition accuracy by 3D-ResNet. TABLE III shows the evaluation results. In addition, Fig. 15 shows the 3D-ResNet accuracy of each person individually.
TABLE II
ANONYMIZATION PERFORMANCE AGAINST GAIT RECOGNITION SYSTEM ENHANCED BY DATA AUGMENTATION

| Method         | Gait recognition accuracy | GEI   | FDF   | SFDEI |
|----------------|---------------------------|-------|-------|-------|
| non-anonymized |                            | 100%  | 100%  | 100%  |
| before-HRTT    | phase-only                | 26.3% | 19.5% | 21.1% |
|                | shape-only                | 11.3% | 9.47% | 13.2% |
|                | both                      | 1.05% | 1.05% | 2.63% |
| after-HRTT     | phase-only                | 31.6% | 23.7% | 23.7% |
|                | shape-only                | 15.3% | 13.2% | 15.8% |
|                | both                      | 5.26% | 7.89% | 5.79% |

Fig. 14. Local shape comparison of before-HRTT and after-HRTT. Original gait frame has a dent in its foot region. This local shape is once removed by silhouette deformation but restored by human region texture transfer (HRTT) in anonymized gait frame, because of too smooth displacement field (DF).

TABLE III
EXPERIMENTAL RESULTS ON PERSON DETECTION ACCURACY BY YOLO AND “WALKING” ACTION RECOGNITION ACCURACY BY 3D-ResNET (VISUAL NATURALNESS)

| Method         | YOLO accuracy | 3D-ResNet accuracy |
|----------------|---------------|--------------------|
| non-anonymized | 100%          | 75.6%              |
| phase-only     | 99.9%         | 70.3%              |
| shape-only     | 100%          | 73.0%              |
| both           | 100%          | 73.0%              |
| pixelization   | 0.00%         | 29.6%              |
| blurring       | 25.7%         | 23.2%              |

For the non-anonymized videos, YOLO achieves a detection accuracy of 100%, whereas 3D-ResNet only achieves a recognition accuracy of 75.6%. This is because action recognition from a video is a more challenging task than person detection in an image. Here, our focus is on how much these accuracies are degraded by the gait anonymization process. If there is no degradation, we can claim that the anonymization process can keep the visual naturalness of an input video.

For all the strategies, i.e., phase-only, shape-only, and both, it can be seen that the YOLO accuracy is not degraded and the 3D-ResNet accuracy is only slightly degraded. These results indicate that the proposed method can keep the visual naturalness, especially in the case of shape-only and both strategies. The anonymized videos of the people with ID:1, ID:4, ID:7, and ID:12 were hardly correctly recognized as “walking” action by 3D-ResNet, which is, however, also the case with their original gait videos.

Pixelization and blurring cannot keep the YOLO accuracy nor the 3D-ResNet accuracy. This demonstrates that visual abstraction-based approaches cannot keep the visual naturalness at all, as mentioned in Section II.

VI. CONCLUSION

In this paper, we propose a method for anonymizing gait information in web videos to reduce the risk of privacy leakage caused by gait recognition systems. The proposed method consists of two modules: silhouette deformation and human region texture transfer. The former slightly deforms the silhouette of each frame in an input gait video from not only the static aspect (i.e., shape) but also the dynamic aspect (i.e., phase) so that the person in the input video cannot be correctly recognized. Anonymization from the dynamic aspect is an important contribution of this paper. The latter, namely HRTT, is achieved by the use of a displacement field in order to handle various types, colors, and designs of clothes without any training process, which is another contribution.

In our experimental results, the proposed method succeeded in reducing the gait recognition accuracy from 100% to at most 12.1% (4.73% in the lowest case) by only static anonymization. More importantly, the accuracy was further reduced to at most 8.42% (1.57% in the lowest case) by the combination of static and dynamic anonymization. This trend can be seen in whatever kind of gait features (i.e., GEI, FDF, and SFDEI) was used for gait recognition. This demonstrates the high anonymization capability of the proposed method.

In addition, the proposed method also succeeded in preserving the visual naturalness of the walking people’s appearance in terms of the action recognition accuracy by 3D-ResNet, which was 75.6% and 73.0% before and after anonymization, respectively. This result indicates that the proposed method does not significantly affect the visual naturalness of the input video. We also verified this fact by the qualitative evaluation.

Importantly, the proposed dynamic anonymization technique (i.e., phase perturbation) can be combined with not only our own static anonymization technique (i.e., shape perturbation) but also any other existing shape-based gait anonymization method, as below. First, the phase perturbation technique is applied to an input gait video (this process is equivalent to the phase-only strategy), whose result is next input into an existing shape-based method. Then, the anonymized gait silhouette is colorized by the proposed texture transfer technique. With this procedure, the phase perturbation...
technique can extend the performance of existing shape-based gait anonymization methods. In the current study, the hyper-parameters of HRTT were empirically given. However, this is not necessarily optimal. In fact, local shapes of the human region before anonymization could be restored by HRTT with a too smooth displacement field, as we experimentally examined. This slightly degrades the anonymization performance of the proposed method. We will improve this problem in our future research. In addition, we will extend the proposed method so that it can cope with the case where people’s walking direction is changed within a single video.

APPENDIX

A. Details of the Phase Value Estimation

We use DP matching, also known as Dynamic Time Warping (DW), to make the best correspondence between \( \mathcal{R} = (R_1, \ldots, R_N) \) and \( \mathcal{S} = (S_1, \ldots, S_M) \).

Suppose a lattice whose size is \( N \times M \), and let \( \mathcal{P} \) be a monotonic increasing path on the lattice starting at \( (1, 1) \) and ending at \( (N, M) \), as shown in the left side of Fig. 16. This \( \mathcal{P} \) gives a correspondence between \( \mathcal{R} \) and \( \mathcal{S} \), whose cost is

\[
\tilde{C}(\mathcal{R}, \mathcal{S}; \mathcal{P}) = \sum_{(i,j) \in \mathcal{P}} \text{cost}(R_i, S_j).
\]

(26)

The frame-wise cost function \( \text{cost}(\cdot, \cdot) \) can be variously designed. In our implementation, we simply design it as

\[
\text{cost}(R_i, S_j) = \sum_{y=1}^{H} \sum_{x=1}^{W} |R_i(x, y) - S_j(x, y)|,
\]

(27)

where \( R_i(x, y) \) and \( S_j(x, y) \) are the pixel value of the silhouette frame \( R_i \) and that of \( S_j \) on the pixel \( (x, y) \). \( H \) and \( W \) are the image height and width. Under the above setting, DP matching allows us to obtain the best path \( \hat{\mathcal{P}} \) that minimizes \( \tilde{C}(\mathcal{R}, \mathcal{S}; \mathcal{P}) \), namely,

\[
\hat{\mathcal{P}} = \arg\min_{\mathcal{P}} \tilde{C}(\mathcal{R}, \mathcal{S}; \mathcal{P}).
\]

(28)

Once \( \hat{\mathcal{P}} \) is obtained, the best matching cost

\[
C(\mathcal{R}, \mathcal{S}) = \min_{\mathcal{P}} \tilde{C}(\mathcal{R}, \mathcal{S}; \mathcal{P}) = \tilde{C}(\mathcal{R}, \mathcal{S}; \hat{\mathcal{P}})
\]

(29)

can be calculated.

We apply the above process to not only \( \mathcal{S} \) but also its periodic shifted version \( \mathcal{S}_l \) for all \( l \), where \( 0 \leq l < M \), as mentioned in Subsection III-B introduced in Subsection III-C.

Finally, for each frame \( S_j \), we construct an index set \( \mathcal{N}_j = \{ i | (i, j) \in \hat{\mathcal{P}}_l \} \) and estimate the phase value of \( S_j \) as

\[
\hat{\varphi}_j = \frac{1}{|\mathcal{N}_j|} \sum_{i \in \mathcal{N}_j} 2\pi \frac{i - 1}{N},
\]

(31)

where \( 2\pi \frac{i - 1}{N} \) is the phase value of \( R_i \).

B. Details of the Training Process of \( E \) and \( D \)

As mentioned in Subsection III-C, we first train a VAE to obtain \( \hat{\mathbf{z}}_i^{\text{gt}} \), whose network structure is shown in Fig. 17. In the figure, “FC” means a fully-connected layer, where \( n \) is the number of units in it. “Conv” and “Deconv” means a convolution layer and a transposed convolution layer, respectively, where “KS” and “Ch” are their kernel size and the number of channels. Hereafter, let \( E_{\text{vae}} \) and \( D_{\text{vae}} \) be the encoder and the decoder parts of the VAE, respectively. We use sigmoid activation after the last layer of \( D_{\text{vae}} \) while do not use any activation after the last layer of \( E_{\text{vae}} \). For the other layers, we use ReLU activation after them.

Using the above \( E_{\text{vae}} \), we extract a feature vector \( \xi_{\hat{\mathbf{z}}_i}^{\text{gt}}(\theta_i) \) from \( \text{Sil}_{\hat{\mathbf{z}}_i}^{\text{gt}}(\theta_i) \) for all \( i \) and obtain \( \hat{\mathbf{z}}_i^{\text{gt}} \) based on Formula (3). After that, using the \( \hat{\mathbf{z}}_i^{\text{gt}} \), we train the \( E \) and \( D \). The network structure of \( E \) is identical to that of \( E_{\text{vae}} \) except that it does not have a fully-connected layer to calculate the standard deviation (std. dev.). The \( E \) and \( E_{\text{vae}} \) share the same structure but their parameters are independently trained from each other. The network structure of \( D \) is shown in Fig. 18. Importantly, we design the decoder \( D \) by cascading a subnetwork \( F \) and the above \( D_{\text{vae}} \), whose parameters are fixed. This means that we only tune the parameters of the subnetwork \( F \) during the training process of \( D \). We previously show the loss function

Fig. 16. DP matching between reference sequence \( \mathcal{R} \) and target sequence \( \mathcal{S} \).

Fig. 17. Network structure of VAE introduced in Subsection III-C.
as Formula (4), but it is a simplified version. We actually use the following loss function, namely,

$$L(E, F)$$

$$= \sum_{\alpha} \sum_{i=1}^{M_{\alpha}} \left( ||E[\text{Sil}_\alpha(\theta_i)] - z_{\alpha i}^{\text{gt}} ||^2 
+ \lambda \left( ||\tilde{\xi}_{\alpha, i}(\theta_i) - \xi_{\alpha, i}(\theta_i) ||^2 + ||D_{\text{vae}}[\tilde{\xi}_{\alpha, i} - \text{Sil}_\alpha(\theta_i)] ||^2 \right) \right),$$

(32)

where $\tilde{\xi}_{\alpha, i} = F[E[\text{Sil}_\alpha(\theta_i)], p_{\theta_0}]$. Thus, the role of the subnetwork $F$ is to reconstruct the VAE-based feature vector $\xi_{\alpha, i}(\theta_i)$ from the shape code $z_{\alpha} \approx E[\text{Sil}_\alpha(\theta_i)] \approx z_{\alpha i}^{\text{gt}}$ and the phase vector $p_{\theta_0}$.

Note that we do not necessarily have to employ Formula (3) to define $z_{\alpha i}^{\text{gt}}$. Definitions other than the average such as the element-wise maximum and minimum can also be employed. However, the average is one of the best choices because of the following reason. Since walking is a periodic action, the vector set $\{\xi_{\alpha}(\theta_i)\}$ forms a closed curve-like continuous manifold in the feature space. The $\xi_{\alpha}(\theta_i)$ calculated above is a point on the manifold. On the other hand, the $z_{\alpha i}^{\text{gt}}$ is a certain representative value of the manifold, which does not necessarily lie on the manifold itself. The situation is depicted in Fig. 19. To enable the $F$ to precisely reconstruct $\xi_{\alpha, i}(\theta_i)$ from $z_{\alpha} \approx z_{\alpha i}^{\text{gt}}$ and $p_{\theta_0}$, the deviation vector $\xi_{\alpha, i}(\theta_i) - z_{\alpha i}^{\text{gt}}$ should be strongly correlated with $p_{\theta_0}$; in other words, the $\xi_{\alpha, i}(\theta_i) - z_{\alpha i}^{\text{gt}}$ is desirable to have a different direction for different $\theta_i$. This property can be obtained when the $z_{\alpha i}^{\text{gt}}$ is near the center of the manifold. Therefore, Formula (3) is one of the best choices.

**REFERENCES**

[1] W. Zeng, C. Wang, and Y. Li, “Model-based human gait recognition via deterministic learning,” *Cognit. Comput.*, vol. 6, no. 2, pp. 218–229, 2014.

[2] Y. Makihara, D. Matovski, M. Nixon, J. Carter, and Y. Yagi, *Gait Recognition: Databases, Representations, and Applications*. Hoboken, NJ, USA: Wiley, 2015, pp. 1–15.

[3] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, “Gait verification system for criminal investigation,” *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, pp. 163–175, Oct. 2013.

[4] D. Muramatsu, Y. Makihara, H. Iwama, T. Tanoue, and Y. Yagi, “Gait verification system for supporting criminal investigation,” in *Proc. 2nd IAPR Asian Conf. Pattern Recognit.*, Nov. 2013, pp. 747–748.

[5] P. Agrawal and P. J. Narayanan, “Person de-identification in videos,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 299–310, Mar. 2011.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[7] P. Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *Proc. 7th Int. Conf. Document Anal. Recognit.*, 2003, pp. 958–962.

[8] Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi, “Anonymization of gait silhouette video by perturbing its phase and shape components,” in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1679–1685.

[9] M. Boyle, C. Edwards, and S. Greenberg, “The effects of filtered video on awareness and privacy,” in *Proc. ACM Conf. Comput. Supported Cooperat. Work (CSCW)*, 2000, pp. 1–10.

[10] C. Neustaedter, S. Greenberg, and M. Boyle, “Blur filtration fails to preserve privacy for home-based video conferencing,” *ACM Trans. Comput.-Hum. Interact.*, vol. 13, no. 1, pp. 1–36, 2006.

[11] E. M. Newton, L. Sweeney, and B. Malin, “Preserving privacy by de-identifying face images,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 223–243, Feb. 2005.

[12] N. Ruchaud and J. L. Dugelay, “Automatic face anonymization in visual data: Are we really well protected?” *Electron. Imag.*, vol. 15, pp. 1–7, Feb. 2016.

[13] Y. Nakashima, T. Ikeno, and N. Babaguchi, “Evaluating protection capability for visual privacy information,” *IEEE Security Privacy*, vol. 14, no. 1, pp. 55–61, Jan. 2016.

[14] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, “Face swapping: Automatically replacing faces in photographs,” *ACM Trans. Graph.*, vol. 27, no. 3, pp. 39.1–39.8, 2008.

[15] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, “Integrating utility into face de-identification,” in *Proc. 5th Int. Conf. Privacy Enhancing Technol.*, 2005, pp. 227–242.

[16] R. Gross, L. Sweeney, F. D. L. Torre, and S. Baker, “Model-based face de-identification,” in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, 2006, p. 161.

[17] L. Sweeney, “k-anonymity: A model for protecting privacy,” *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[18] Y. Nakashima, T. Koyama, N. Yokoya, and N. Babaguchi, “Facial expression preserving privacy protection using image melding,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6.

[19] I. Goodfellow et al., “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.

[20] K. Brkic, I. Sikiric, T. Hekac, and Z. Kalafatic, “I know that person: Generative full body and face de-identification of people in images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1319–1328.

[21] Z. Ren, Y. J. Lee, and M. S. Ryoo, “Learning to anonymize faces for privacy preserving action detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 620–636.

[22] O. Gafi, L. Wolf, and Y. Taigman, “Live face de-identification in videos,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9377–9386.

[23] H. Hukkelas, R. Mester, and F. Lindseth, “DeepPrivacy: A generative adversarial network for face anonymization,” in *Proc. Int. Symp. Vis. Comput.*, 2019, pp. 565–578.

[24] I. Mitsumagi, M. Mukanoki, Y. Kawanishi, H. Hattori, and M. Minoh, “Privacy-protected camera for the sensing web,” in *Proc. 13th Int. Conf. Inf. Process. Manag. Uncertainty Knowl.-Based Syst.*, 2010, pp. 622–631.

[25] N.-D.-T. Tieu, H. H. Nguyen, H.-Q. Nguyen-Son, J. Yamagishi, and I. Echizen, “An approach for gait anonymization using deep learning,” in *Proc. IEEE Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2017, pp. 1–6.
[26] N.-D.-T. Tieu, H. H. Nguyen, H.-Q. Nguyen-Son, J. Yamagishi, and I. Echizen, “Spatio-temporal generative adversarial network for gait anonymization,” J. Inf. Secur. Appl., vol. 46, pp. 307–319, Jun. 2019.

[27] N.-D.-T. Tieu, H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “An RGB gait anonymization model for low-quality silhouettes,” in Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Nov. 2019, pp. 1686–1693.

[28] L. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, arXiv:1412.6572.

[29] Z. He, W. Wang, J. Dong, and T. Tan, “Temporal sparse adversarial attack on sequence-based gait recognition,” 2020, arXiv:2002.09674.

[30] M. Maqsood, M. A. Ghazanfar, I. Mehmood, E. Hwang, and S. Rho, “A meta-heuristic optimization based less imperceptible adversarial attack on gait based surveillance systems,” I. Signal Process. Syst., pp. 1–23, Feb. 2022. doi: 10.1007/s11265-022-01742-x.

[31] A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger, “Gait-based recognition of humans using continuous HMMs,” in Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit., May 2002, pp. 336–341.

[32] J. Man and B. Bhanu, “Individual recognition using gait energy image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 2, pp. 316–322, Feb. 2006.

[33] K. Bashir, T. Xiang, and S. Gong, “Gait recognition using gait entropy image,” in Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP), 2009, pp. 1–6.

[34] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, “Gait recognition using a view transformation model in the frequency domain,” in Proc. Eur. Conf. Comput. Vis., 2006, pp. 151–163.

[35] N. Setoguchi, K. Nakashima, W. Tong, Y. Iwashita, and R. Kurazume, “A CNN-based gait recognition robust to low resolution images using inter-image difference,” in Proc. 14th Joint Workshop Mach. Perception Robot., 2018.

[36] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “GEINet: View-invariant gait recognition using a convolutional neural network,” in Proc. Int. Conf. Biometrics (ICB), Jun. 2016, pp. 1–8.

[37] 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 Trans. Pattern Anal. Mach. Intell., vol. 39, no. 2, pp. 209–226, Mar. 2017.

[38] Z. Zhu et al., “Gait recognition in the wild: A benchmark,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2021, pp. 14789–14799.

[39] S. Zhang, Y. Wang, and A. Li, “Cross-view gait recognition with deep universal linear embeddings,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 9095–9104.

[40] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, “Disentangled person image generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 99–108.

[41] O. Gafni, O. Ashual, and L. Wolf, “Single-shot freestyle dance reenactment,” 2020, arXiv:2012.01158.

[42] Y. Makihara et al., “The OU-ISIR gait database comprising the treadmill dataset,” IPSJ Trans. Comput. Vis. Appl., vol. 4, pp. 53–62, Apr. 2012.

[43] K. Hara, H. Katoaka, and Y. Satoh, “Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6546–6555.

[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

Kazuaki Nakamura (Member, IEEE) received the B.S. degree in engineering and the M.S. and Ph.D. degrees in informatics from Kyoto University in 2005, 2007, and 2011, respectively. From 2012 to 2021, he was an Assistant Professor with the Graduate School of Engineering, Osaka University. He is currently an Associate Professor with the Faculty of Engineering, Tokyo University of Science. His research interests include image processing, image recognition, video analysis, and multimedia security. He is a member of IEICE, IPSJ, and ITEE.

Naoko Nitta (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in engineering from Osaka University, Japan, in 1998, 2000, and 2003, respectively. From 2002 to 2004, she was a Research Fellow of the Japan Society for the Promotion of Science. From 2003 to 2004, she was a Visiting Scholar at Columbia University, New York. From 2004 to 2022, she was an Assistant Professor, a Lecturer, and an Associate Professor at the Graduate School of Engineering, Osaka University. She is currently a Professor at the Department of Informatics and Mediology, School of Human Environmental Sciences, Mukogawa Women’s University, Japan. Her research interests are in the areas of content analysis of multimedia data, such as videos, images, and social media and in the areas of real-world sensing.

Noboru Babaguchi (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Japan, in 1979, 1981, and 1984, respectively. He is currently an Emeritus Professor and specially appointed as a Professor at Osaka University. He is also a Professor with the Fukui University of Technology. He has published over 250 journals and conference papers and several textbooks. His research interests include image/video analysis, multimedia computing, and intelligent systems.

Yuki Hirose received the B.E. and M.E. degrees in engineering science from Osaka University in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the Graduate School of Engineering. His research interests include gait anonymization and face image generation. He is a Student Member of IEICE.

He has recently been researching privacy protection methods for visual information and the security and fabrication of multimedia. He is a fellow of the IEICE and a member of the ACM, ITE, IPSJ, and JSIA. He also served as the General Co-Chair of MMM2008, the General Co-Chair of ACM Multimedia 2012, the Track Co-Chair of ICPR2012, the Area Chair of IEEE ICME2015, and the Honorary Co-Chair of ACM ICMR2018. He served as a Guest Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, Special Issue on Intelligent Video Surveillance for Public Security & Personal Privacy.