Multi-Modal Human Authentication Using Silhouettes, Gait and RGB

Yuxiang Guo*, Cheng Peng*, Chun Pong Lau, Rama Chellappa

Johns Hopkins University, Baltimore, MD, USA

2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-4544-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/FG57933.2023.10042572

***Abstract*— Whole-body-based human authentication is a promising approach for remote biometrics scenarios. Current literature focuses on either body recognition based on RGB images or gait recognition based on body shapes and walking patterns; both have their advantages and drawbacks. In this work, we propose Dual-Modal Ensemble (DME), which combines both RGB and silhouette data to achieve more robust performances for indoor and outdoor whole-body based recognition. Within DME, we propose GaitPattern, which is inspired by the double helical gait pattern used in traditional gait analysis. The GaitPattern contributes to robust identification performance over a large range of viewing angles. Extensive experimental results on the CASIA-B dataset demonstrate that the proposed method outperforms state-of-the-art recognition systems. We also provide experimental results using the newly collected BRIAR dataset.**

I. Introduction

Body recognition from videos is an important yet challenging computer vision task. The objective is to determine whether the subjects in different videos have the same identity. Similar to face recognition, body recognition has applications ranging from surveillance to intelligent transportation. In particular, body recognition is more robust than face recognition in many unconstrained situations, where faces are acquired in non-cooperative conditions. Gait recognition [6], [18], [5], which recognizes people based on their walking patterns over time, similarly performs recognition based on the body, but typically uses human silhouettes as input. In comparison, body recognition methods exploit silhouettes as well as RGB data.

There are advantages and disadvantages when body or gait is used for remote identification separately. Specifically, RGB videos contain rich information that can lead to significantly better performance; however, such information can be ineffective in situations involving clothing change and image quality degradation. The silhouette modality allows a recognition network to focus purely on the subject's body shape and gait; as such it is less susceptible to clothing changes. This robustness comes at the cost of generally lower performance compared to methods that exploit RGB data. Motivated by this observation, we propose Dual-Modal Ensemble (DME), which performs learning in both RGB and silhouette domains and leverages the model ensemble to extract the most robust features for unconstrained situations. DME is also flexible and can be used separately when the image condition is ill-suited for an individual modality.

*Equal contribution.

979-8-3503-4544-5/23/$31.00 ©2023 IEEE

Furthermore, within DME, we address several issues in the *gait recognition* space. Due to the high computational cost of processing video data through neural networks, State-of-The-Art (SoTA) methods rely on temporal pooling operations to reduce the dimension from 3D to 2D. As such, spatial information plays a major role in these recognition algorithms. It has been observed across the gait recognition literature that such systems tend to generalize poorly over different view angles [6], [18], [5]. An efficient gait representation in the temporal domain is thus highly desired and can complement conventional gait recognition systems. Such a temporal representation can provide knowledge of the camera's viewing angle, assuming the cameras are stationary and the subject is walking in a normal pattern.

Inspired by traditional gait recognition approaches that produce such a temporal representation through a Double Helical Signature (DHS) [24], we propose GaitPattern, learned using a deep neural network and incorporate it into the recognition system. We empirically find that the introduction of GaitPattern significantly improves performances at challenging view angles.

In summary, this paper makes the following contributions:

- We propose Dual-Modal Ensemble, which incorporates both gait and RGB modalities to perform body recognition; such a design allows for both flexibility and superior performance both for indoor and outdoor scenes.
- We propose GaitPattern, which efficiently incorporates temporal information into gait recognition and leads to performance improvements in indoor scenes, especially for challenging viewing angles.
- We perform extensive evaluations on both CASIA-B and Biometric Recognition and Identification at Altitude and Range (BRIAR) dataset, an unconstrained face and body recognition dataset.

II. Related Work

A. Gait Recognition

Gait recognition aims to identify human subjects by their walking pattern. In general, gait representation can be done in 2D and 3D. There are many works [34], [4], [2] that use 3D gait representations, which are obtained from multi-camera setups. While 3D gait representation contains rich information and provides compelling performances, a multi-camera setup is often impractical for general applications. Moreover, such a 3D approach is computationaly costlier than 2D approaches. There are typically three approaches

to obtain 2D gait representation. One approach uses gait silhouettes [6], [18], [5], [10], [29], [11], [12], [13], [21], [8], extracted from video frames of walking subjects. The skeleton-based approaches [1], [17], [26], [25] use the key-points of the joints, i.e. a skeleton representation, and extract features from the skeleton. Another approach fuses features from silhouettes and skeleton [27], [23].

B. Double Helical Signature

Human gait represented as a Double Helical Signature (DHS) [24], also known as a type of "Frieze pattern" [22], has been proposed and applied for gait sequence analysis in the past, when limited computational power encouraged researchers to find efficient representations to perform recognition. As shown in Fig. 2, DHS can be generated by extracting the horizontal slices of the subject's knee from every frame in a video and stacking the slices to form a time-width image. Prior works [24], [22], [19], [20] have shown that DHS can efficiently encode parameters such as step size and gait/walking rate and also determine if the subject is carrying objects.

C. Video-based Body Recognition

Due to the availability of rich spacial and temporal information and computational resources, videos can be more effective for whole body recognition. Accurate capture of temporal information is a primary challenge. There are mainly two methods to capture the temporal relations. Temporal attention [7], [38], [35] is one method, which aims to determine the effectiveness of each frame and discards the low-quality ones. The other method is self-attention or Graph Convolutional Networks (GCNs) [16], [30], [3], which exploits the temporal relations.

D. Remote Biometric Recognition in the Wild

Remote biometrics in the wild has been studied for over fifteen years. Some of the early efforts in remote biometrics recognition were based on gait and faces. As discussed before, the gait recognition works reported in [8], [12], [13], [21] and video-based recognition works reported in [36], [37] are some examples of early works on remote biometric recognition efforts, although the images and videos were collected at varying degrees of complexity. More recently, face recognition at distances of 300-1000 meters has been addressed in [15], [14], [31], [32]. For example, [15], [14] propose a generative single frame restoration algorithm which disentangles the blur and deformation due to atmospheric turbulence and reconstructs a restored image. Extensive experiments demonstrate the effectiveness of the proposed restoration algorithm, which achieves satisfactory results at 300 and 650 meters, but the low-accuracy detection of faces at 1000 meters affects the overall performance. Similar efforts using generative models have been reported in [31], [32]. While many datasets such as CASIA-B and CASIA-E have been collected and many traditional and deep learning-based algorithms have been evaluated on these datasets, these datasets also vary in terms of how unconstrained the collected videos are. Recently, the IARPA BRIAR program has collected datasets of faces and whole bodies at distances upto 500 meters. We present the results of whole-body recognition based on silhouettes, gait and RGB using the BRIAR dataset.

III. Method

A. Problem Formulation

Suppose the target video $\boldsymbol{V} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T] \in \mathcal{F} := \mathbb{R}^{H\times W\times T\times 3}$ consists of T video frames $\boldsymbol{f}_i$, where H and W are the height and width of the frame. In $\boldsymbol{V}$, we assume there is a corresponding subject, labeled as $y \in \mathcal{Y}$, where $y \in \{1, 2, \ldots, |\mathcal{Y}|\}$ and $\mathcal{Y}$ is the dataset. To represent gait, we obtain the human silhouette sequence $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T] \in \mathcal{S} := \{0, 1\}^{H\times W\times T\times 1}$, which consists of binary masks obtained by image segmentation. Let F_θ be a parameterized model which maps any video in $\mathcal{F}$ to a feature vector $\boldsymbol{X}$ in $F_\theta(\mathcal{Y})$. An accurate whole-body recognition system can map two videos $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ with the same identity to features that are close in a certain feature space, i.e. $\mathcal{D}(F_\theta(\boldsymbol{V}_1), F_\theta(\boldsymbol{V}_2))$, where $\mathcal{D}$ is an appropriate distance function, e.g. Euclidean. Specifically, this can be done by using either RGB frames, silhouette sequences, or both. For body recognition using RGB frames, the videos are first preprocessed by masking out the background based one $\boldsymbol{S}$ to prevent overfitting, i.e. $\boldsymbol{V} = \boldsymbol{V}_{\text{orig}} \odot \boldsymbol{S}$, where $\boldsymbol{V}_{\text{orig}}$ is the original video and $\odot$ is the Hadamard product.

In the following sections, we describe DME, which uses RGB and silhouette together for more robust performances, and *GaitPattern*, a gait recognition method that leverages the efficient Double Helical Signature representation [24].

B. Dual-Modal Ensemble

Dual-Modal Ensemble proposes to use two CNN architectures to extract features from individual modalities. As shown in Fig. 1, this process can be expressed as:

$$X_f = \mathcal{F}_f(\boldsymbol{V}), X_s = \mathcal{F}_s(\boldsymbol{S}), \tag{1}$$

where $\mathcal{F}_f$ and $\mathcal{F}_s$ are the feature extraction CNNs for the respective modalities. Generally, $\mathcal{F}_f$ and $\mathcal{F}_s$ can be any well-performing CNN architecture, and we elaborate the details of network design in Section III-C and III-D. Using separate feature extraction blocks for each modality allows for better flexibility, as compared to combining the inputs and feature extraction stages in a monolithic framework. Specifically, if one input mode suffers from obviously degraded quality, DME can still use the other mode to perform recognition. We highlight this flexibility in Section IV-C.

After X_f and X_s are extracted, they pass through the respective MLP networks to extract the video identification embeddings l_f and l_s, which are defined as,

$$l_f = \mathcal{M}_f(X_f), l_s = \mathcal{M}_s(X_s), \tag{2}$$

where $\mathcal{M}_f$ and $\mathcal{M}_s$ are the MLP networks. For identification, a triplet loss [9] is used to maximize the distance

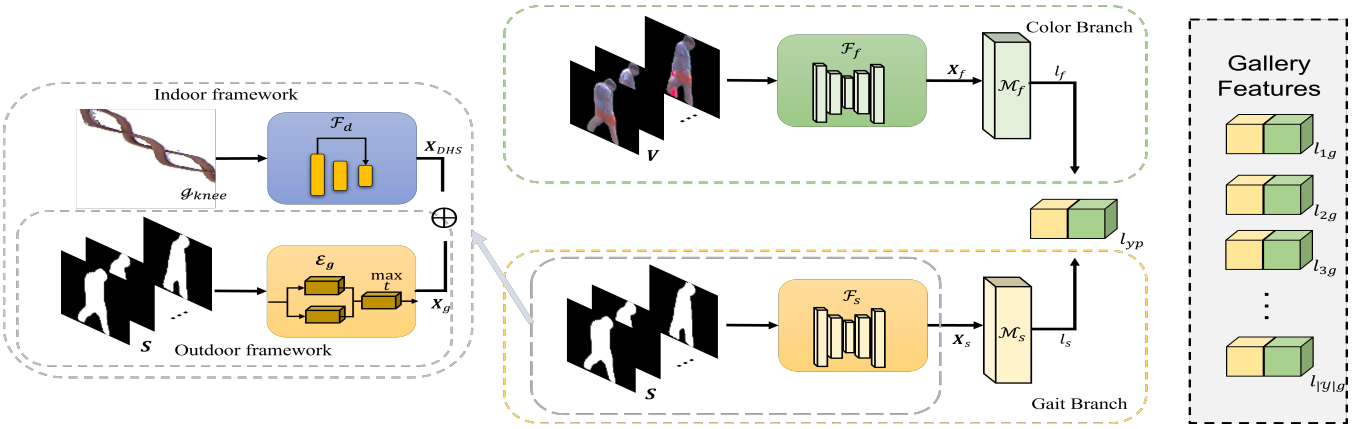


Fig. 1: The overall pipeline of the proposed method. The method is divided into gait and RGB two branches. They both use backbone convolutions $\mathcal{F}_f$ and $\mathcal{F}_s$ following the multiple-layer perceptrons (MLP) $\mathcal{M}_f$ and $\mathcal{M}_s$ to generate the feature l_y. The features are extracted from the gallery (l_{yg}) and probe (l_{yp}) videos to measure the distance between them. For gait branch, the DHS feature X_{DHS} will be concatenated with the that of gait X_g, if the DHS feature is available.

of embeddings from different subjects, and minimize the distance from the same subject. Specifically:

$$\begin{aligned}\mathcal{L}_f^{tri} &= [\mathcal{D}(l_f(i), l_f(k)) - \mathcal{D}(l_f(i), l_f(j)) + m]_+, \\ \mathcal{L}_s^{tri} &= [\mathcal{D}(l_s(i), l_s(k)) - \mathcal{D}(l_s(i), l_s(j)) + m]_+,\end{aligned} \tag{3}$$

where i and j are videos from the same subject, k is the video from another subject. $\mathcal{D}(*)$ is the Euclidean distance measure, m is the margin of the triplet loss, and operation $[*]_+$ describes $\max(*, 0)$. The overall loss is:

$$\mathcal{L}^{tot} = \mathcal{L}_f^{tri} + \mathcal{L}_s^{tri}. \tag{4}$$

At test time, DME obtains the *ensemble feature* through concatenation, i.e. $l_i = l_f \oplus l_s$ from every video, and uses l_i to perform recognition. We find this to be a simple yet effective strategy that produces more discriminative features than those from individual modalities.

In the following section, we expand on the design details of individual feature extraction components; in particular, we propose a novel GaitPattern framework.

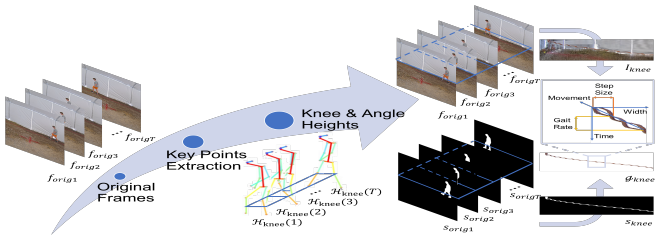


Fig. 2: The process of DHS generation.

C. GaitPattern

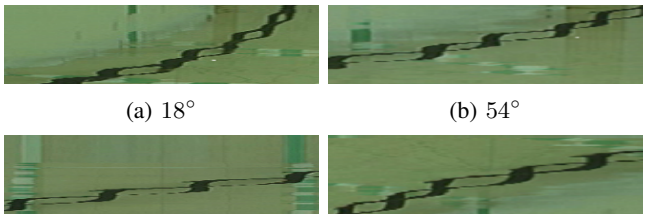
A gait representation framework should focus on the walking motion and body shape and generate features robust to the identities' appearance. We propose GaitPattern, which incorporates both conventional silhouette-based feature extractor and an additional novel Double Helical Signature branch to better encode information like motion and viewing angles.

1) Gait Feature Extraction: A standard gait feature extraction model takes the silhouette sequences $\boldsymbol{S}$ as input and generates the gait feature X_g. Previous works [6], [5], [18] mainly apply the "CL-SP-CL-SP-CL" pattern feature extractor in which CL represents a convolutional layer and SP, a spatial pooling layer. They use different kinds of convolution, i.e. 2D and 3D convolution, etc., and structures, i.e., global, local features combination and feature map partition, etc., to generate features. Then a temporal pooling step is applied to shrink the time dimension to extract a same-sized feature. This process is described as follows:

$$X_g = \mathcal{F}_s(S) = \max_t(\mathcal{E}_g(\boldsymbol{S})), \tag{5}$$

where $\mathcal{E}_g$ is a gait feature extractor and $\max_t$ represents temporal pooling. Our gait feature extraction model could flexibly apply the recent gait recognition methods as a backbone. By introducing the temporal pooling operation $\max_t$, the temporal information is lost, so we propose a different method to increase the temporal resolution.

2) DHS Feature Extraction: Based on previous works, it appears that gait recognition performance is reduced due to information loss caused by temporal pooling and different viewing angles. Based on our analysis, we desire to preserve some information in the time domain to make up for the information loss caused by temporal pooling. Then it is crucial to provide viewpoint information and regularize the model to generate robust features. We observe that the DHS contains this information as it records the dynamic movement of knees in 2D. The knee movement is representative and a distinguishable feature of human gait. The DHS patterns as a function of viewing angles are shown in Fig. 3. We embed the viewing angle information into the DHS to help the model generate robust features for the same identity in different angles, which mitigates the effects of feature variations at different view points.



(a) 18° (b) 54° (c) 90° (d) 104°

Fig. 3: The DHS generated for CASIA-B at various viewing angles.

As mentioned in II-B, DHS is an efficient representation that describes many gait attributes. In the past, DHS had been estimated at fixed view points and known location of the subjects' knees. We improve this feature extraction process with a CNN-based key-point extraction model and obtain the knee heights $\mathcal{H}_{\text{knee}}(t) \in \mathbb{R}$ in every frame $\boldsymbol{f}_i$. Along with the silhouette mask $\boldsymbol{s}_i$, this allows us to obtain the RGB width-time image $\boldsymbol{I}_{\text{knee}}$, its mask $\boldsymbol{s}_{\text{knee}}$, and subsequently the DHS g_{DHS}. As shown in Fig. 2, this process can be described as:

$$\begin{aligned} \boldsymbol{I}_{\text{knee}}(x,t) &= \boldsymbol{f}_t(x, \mathcal{H}_{\text{knee}}(t)), \\ \boldsymbol{s}_{\text{knee}}(x,t) &= \boldsymbol{s}_t(x, \mathcal{H}_{\text{knee}}(t)), \\ g_{\text{DHS}} &= \boldsymbol{I}_{\text{knee}} \odot \boldsymbol{s}_{\text{knee}}. \end{aligned} \tag{6}$$

After the DHS is generated, we extract the appropriate features from it and incorporate into a modern gait recognition architecture. In this process, one key consideration is the duration of the video. Since we want to increase the temporal resolution, we do not apply temporal pooling which is widely used to deal with varying temporal length. Based on the Ground Principle of Gait Temporal Aggregation [6], to make the system robust, the feature from the sequence containing at least a complete cycle should be the same for periodic motion under ideal conditions. Denote $g_{\text{DHS}}(x, 1:m)$ as the DHS stacked from frame $\boldsymbol{f}_1$ to $\boldsymbol{f}_m$. So the feature extractor for DHS should satisfy the following condition:

$$\mathcal{F}_d(g_{\text{DHS}}(x, s:e)) = \mathcal{F}_d(g_{\text{DHS}}(x, s:s+C)), \quad \forall e-s > C \tag{7}$$

where $\mathcal{F}_d$ is the sequence feature extraction module, C is the complete cycle of the motion, s and e are selected start and end time (height in g_{DHS}), $s, e \in \{1, 2, \ldots, m\}$.

To use the previous principle, we apply $\max(\cdot)$ to select the DHS feature to be fused latter. Therefore, the DHS feature extraction process is designed as follows:

$$\begin{aligned} \boldsymbol{X}_{DHS} = \max(&\mathcal{F}_d(g_{\text{DHS}}(x, s_1:e_1)), \\ &\ldots, \\ &\mathcal{F}_d(g_{\text{DHS}}(x, s_n:e_n))). \end{aligned} \tag{8}$$

So the whole DHS is initially split into same-height intervals and fed to the sequence feature extraction module F_d. The final DHS feature $\boldsymbol{X}_{DHS}$ is the largest one among all intervals' features. Compared to temporal pooling, the DHS feature extraction has the whole view of a video and maintains the temporal resolution with $e-s$ length.

We extract the DHS feature only when the videos are taken in a consistent environment. Since the key-point estimation is accurate in a constrained situation, it makes the extracted DHS reliable. Therefore, for the indoor case, we just adjust the shape of the DHS feature to fit the size of the one from any gait recognition backbones, generating the silhouette modality feature and for outdoor, we directly apply the gait feature. The silhouette feature X_s is as following:

$$X_s = \begin{cases} X_g \oplus X_{DHS}, & \text{indoor}, \\ X_g, & \text{outdoor}. \end{cases} \tag{9}$$

For unconstrained scenarios, only the silhouette feature is used. The silhouette feature X_s is robust to the appearance change, improving the robustness of ensemble feature l_i.

So the GaitPattern is a flexible module which can shift among various gait recognition frameworks, and improves the overall recognition performance across viewing angles.

D. RGB Feature Extraction

For RGB feature extraction, we use a feature extractor similar to gait's feature extractor, but use the video $\boldsymbol{V}$ as the input, i.e. the RGB feature X_f is

$$X_f = (\max_t(\mathcal{E}_f(\boldsymbol{V}))), \tag{10}$$

in which $\mathcal{E}_f$ is a RGB feature extractor. RGB features X_f provide rich information to enhance the appearance representation, and the feature backbones from [6], [5], [18] have shown to be efficient and effective architectures for video feature extraction.

IV. Experiments

A. Datasets

To evaluate the performance of our method, we perform extensive experiments on both controlled and unconstrained datasets. Specifically, we test on CASIA-B [33], which is acquired in a controlled, indoor environment. To further evaluate the robustness of whole body recognition in unconstrained scenarios, we test on the BRIAR dataset.

CASIA-B [33] composes of 124 subjects, each of which has ten recorded sequences. In these ten sequences, six involve normal walking (NM), two involve walking with a bag (BG), and the rest are walking in a coat or jacket (CL). For each sequence, there are eleven videos recorded from viewing points ranging from 0° to 180° at 18° interval. All the videos are recorded in a constrained manner, which means that the videos are recorded indoor with consistent quality. The dataset is divided into training and testing sets following the popular protocol described in [29]: first 74 identities are used for training and rest are for testing. During the training process, all sequences in the training set are fed to the model, i.e. NM, BG, and CL . For testing, the first four NM conditioned sequences (NM#01-NM#04) serve as gallery and the remaining sequence serve as the three probes, i.e. NM#05-NM#06, BG#01-BG#02 and CL#01-CL#02. This is to evaluate the robustness of the system on different walking types and dressing conditions.

Gallery NM#1-4		0° – 180°											
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	mean
NM#5-6	GaitSet [5]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [6]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitGL [18]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitPattern (w/ [5])	92.2	98.4	98.5	97.9	94.7	92.5	95.6	97.1	98.3	98.2	88.4	95.6
	GaitPattern (w/ [6])	96.9	98.4	99.1	98.8	98.7	97.1	98.1	99.0	99.4	99.0	97.3	98.3
	GaitPattern (w/ [18])	96.1	98.7	99.0	98.1	95.6	96.5	97.6	98.3	98.8	98.7	95.1	97.5
	DME (w/ [6])	**99.2**	99.0	99.2	98.9	**99.2**	98.6	99.0	99.3	99.4	99.2	99.4	99.1
	DME (w/ [18])	99.1	**99.2**	**99.2**	**99.1**	99.1	**98.9**	**99.1**	**99.4**	**99.4**	**99.4**	**99.4**	**99.2**
BG#1-2	GaitSet [5]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [6]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitGL [18]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitPattern (w/ [5])	87.2	93.3	95.2	94.2	88.0	82.5	85.8	92.8	96.7	95.4	85.2	90.6
	GaitPattern (w/ [6])	94.2	97.2	97.7	96.9	93.6	91.9	94.7	96.1	98.1	97.4	94.0	95.6
	GaitPattern (w/ [18])	92.8	97.2	98.4	96.7	91.4	90.9	95.0	95.5	97.3	97.5	92.8	95.0
	DME (w/ [6])	98.8	99.0	**99.0**	98.8	97.9	97.2	98.1	99.0	**99.4**	99.2	98.9	98.7
	DME (w/ [18])	**99.2**	**99.2**	98.9	**98.9**	**98.5**	**98.5**	**98.9**	**99.3**	99.3	**99.4**	**99.2**	**99.0**
CL#1-2	GaitSet [5]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [6]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitGL [18]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitPattern (w/ [5])	70.2	84.3	84.5	81.3	74.5	72.5	74.0	79.0	83.0	80.2	66.5	77.3
	GaitPattern (w/ [6])	85.0	92.5	93.4	92.5	88.9	83.7	87.0	90.7	93.0	90.5	87.6	89.5
	GaitPattern (w/ [18])	83.4	93.0	94.6	90.8	85.4	85.0	87.0	89.8	90.9	89.9	84.3	88.5
	DME (w/ [6])	**92.3**	95.9	95.9	94.3	**92.2**	88.5	88.3	93.2	93.7	90.8	86.6	92.0
	DME (w/ [18])	92.1	**96.0**	**96.6**	**95.0**	91.1	**90.3**	**92.0**	**93.2**	**93.8**	**92.6**	**88.7**	**92.8**

TABLE I: Rank-1 accuracy (%) on CASIA-B excluding identical-view case.

BRIAR To evaluate the effectiveness of the proposed approach in unconstrained situations, we experimented with the BRIAR dataset. It includes 158 and 85 subjects serving as training and testing sets. Each identity has indoor and outdoor sequences, namely controlled and field sets. In the controlled set, there are ten cameras simultaneously recording the walking, and there are two walking types, i.e. structure walking and random walking. As for the field set, the videos are recorded at varying distances, i.e. close range, 100m, 200m, 400m, 500m and unmanned aerial vehicle (UAV). Except for the close range set, there is one viewing angle for each distance. And three cameras are applied at close range set in the same position but different yaw angles, i.e. $0°$, $30°$ and $50°$. For all sequences, there are two clothing settings, i.e. set1 and set2 and in addition to the two walking types in controlled set, there are standing videos. Some examples are shown in Fig. 4. For each video, the duration is about 90 seconds. And we get corresponding silhouettes using Detectron2 [28].

We do the unconstrained gait recognition using the following protocol: In the training stage, the videos containing random and structure walking in the training set are fed to the model. In the test stage, we use 'set1' videos in the controlled set as a gallery and 'set2' videos containing gait, i.e. structure and random walking, in the field set as a probe. Typical examples of clothing changes are shown in Fig. 4 second and third columns. The length of the test videos ranges from 5 to 15 seconds.

B. Implementation Details

All the implementations are in Pytorch using four Nvidia A5000 GPUs. We adopt the popular preprocessing method [5] to extract the gait silhouettes and RGB body chip for CASIA-B and BRIAR. The size for each frame is 64×44.

Fig. 4: Examples of a subject in different conditions from the BRIAR dataset at different distances and clothing.The columns represents controlled, 100m-set1, 100m-set2, close range, 200m, 400m, 500m and UAV respectively.

To demonstrate the effectiveness of the proposed modules, GaitPattern and Dual-Model Ensemble, for CASIA-B, we follow the same settings as GaitSet [5], GaitPart [6] and GaitGL [18] for gait and body feature extraction. We use Resnet18 as the DHS feature extraction backbone. For the BRIAR dataset, on account of the complexity and duration of the videos, we enlarge the total iterations to 120k for three models with (8,16) for the batch size, where the first digit is the number of subjects per batch and the second one is the number of videos per subject. All the backbones are trained with randomly sampled thirty frames from a video in each iteration and the selected 5 to 15 seconds sequences are used in the test phase.

Probe	Close range	100m	200m	400m	500m	UAV	Mean
GaitSet	55.40	56.13	30.47	46.76	44.98	65.52	49.87
GaitPart	60.56	71.09	35.62	50.78	50.19	**75.86**	57.35
GaitGL	55.40	66.05	16.31	38.93	40.89	62.07	46.60
RGB (w/ [5])	50.23	57.65	37.34	48.77	39.78	37.93	45.26
RGB (w/ [6])	48.04	45.71	19.96	35.12	35.32	44.83	38.16
RGB (w/ [18])	39.75	45.55	18.03	31.77	24.16	24.14	30.56
DME (w/ [5])	67.92	71.76	**44.21**	**64.65**	56.13	51.72	59.39
DME (w/ [6])	**69.64**	**77.31**	43.35	64.21	**58.36**	72.41	**64.21**
DME (w/ [18])	43.04	51.43	19.96	36.91	27.88	27.59	34.46

TABLE II: Rank-1 accuracy (%) on BRIAR.

C. Quantitative Evaluation

Evaluation on CASIA-B [33] We compare the proposed approach using GaitSet [5], GaitPart [6] and GaitGL [18] with their own original methods on CASIA-B. The Rank-1 accuracy (%) is shown in TABLE I. We see that the GaitPattern improves the overall accuracy and with lower variance. The DME module boost the performance to a higher level.

Compared to the three backbones, the GaitPattern-based method improves the mean accuracy with lower variance across viewing angles, especially for the CL condition, improving by 6.9%, 10.8% and 4.9%, respectively. The fusion of GaitPattern and GaitPart improves performance by 2.1%, 4.1%, 10.8% in NM, BG and CL conditions respectively, reaching 98.3%, 95.6% and 89.5%.

When it comes to DME, we combine the features from silhouettes, DHS and RGB. With the rich information from RGB, there is still a big improvement from the GaitPattern-based methods. And DME with GaitGL as backbone outperforms the others. Compared to the original GaitGL, the increase is 1.8%, 4.5% and 9.2% for NM, BG and CL conditions. Even compared to the GaitPattern-based GaitGL, DME with GaitGL has 1.7%, 4.0% and 4.3% improvements.

Evaluation on BRIAR We also apply GaitSet [5], GaitPart [6] and GaitGL [18] to the proposed method. The Rank-1 accuracy (%) is shown in TABLE II. DME with GaitPart achieves the best overall recognition performance, reaching 64.21%. And GaitSet and GaitPart both have performance improvements of 15.07%, and 6.86% respectively with the help of DME. But their performance have significant increase compared to only applying the RGB branch by 14.13%, 26.05% and 3.85%, respectively. This indicates that the silhouette features enhances the effectiveness of aggregated features. We observe that DME with GaitGL has relatively low performance as the 3D convolution module does not extract robust features from a long video, especially with temporal stride. Further, from the fifth column of Fig. 4, we find that as only the part of body above the knee is observable, the gait recognition performance at 200m is lower than at other distances.

To demonstrate the effectiveness of the RGB feature, we experiment with the clothing protocol. For the BRIAR dataset, in the test stage, we use 'set1' and 'set2' sequences in the controlled set as a gallery and similarly 'set1' and 'set2' sequences in the field set as a probe.

From TABLE III, we see that silhouette features do not perform well due to information loss. But when we introduce DME, the RGB feature helps improve the embedding's performance. For all three models, the DME results are even better than the RGB ones, which shows that the dual modalities complement mutually. Fig. 5 demonstrates this property using distances between a gallery and probes.

	(a) Gallery	(b) Probe	(c) Probe	(d) Probe
Gait		0.84	0.85	1.18
RGB		0.75	0.98	1.05
Dual		0.81	0.91	1.11

Fig. 5: The Euclidean distances of probes to gallery for gait, RGB and dual features (GaitPart backbone).

Backbone	Gait	RGB	Dual
GaitSet [5]	86.31	92.53	95.89
GaitPart [6]	73.29	88.27	94.19
GaitGL [18]	65.56	86.29	87.36

TABLE III: Mean accuracy (%) on the privately collected unconstrained outdoor dataset with changing clothes.

V. Conclusion

In this work, we present Dual-Modal Ensemble, which leverages both the human silhouette and RGB to perform whole-body recognition. Specifically, DME leverages the rich information in RGB image to boost the general recognition performance, while also maintaining high performance on challenging conditions, e.g. with changing clothes, based on a gait recognition sub-network. DME accomplishes this by training a powerful gait and body recognition network individually, and concatenating features produced by each modality to obtain the final feature of the video. DME shows very robust performance in unconstrained conditions, demonstrating the effectiveness of combining features from different modalities. Within DME, we propose a gait recognition network called GaitPattern, inspired by the traditional gait pattern analysis. GaitPattern generates and incorporates a gait representation called Double Helical Signature, and can be used alongside popular gait recognition architectures to complement the lack of view angle information and spatial resolution. We find that GaitPattern can significantly enhance performances of current SoTA methods, particularly over challenging view angles. Future work includes investigating more robust key-point detection algorithm such that GaitPattern can be better incorporated into unconstrained environments, as well as reducing the computational complexity of multi-modal ensemble by fusing models at early stages without losing discriminative power for each modality.

VI. ACKNOWLEDGEMENT

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE transactions on biometrics, behavior, and identity science*, 2(4):421–430, 2020.

[2] G. Ariyanto and M. S. Nixon. Model-based 3d gait biometrics. In *2011 international joint conference on biometrics (IJCB)*, pages 1–7. IEEE, 2011.

[3] S. Bai, B. Ma, H. Chang, R. Huang, and X. Chen. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7339–7348, 2022.

[4] R. Bodor, A. Drenner, D. Fehr, O. Masoud, and N. Papanikolopoulos. View-independent human motion classification using image-based reconstruction. *Image and Vision Computing*, 27(8):1194–1206, 2009.

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

[6] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

[7] Y. Fu, X. Wang, Y. Wei, and T. Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8287–8294, 2019.

[8] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005.

[9] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[10] S. Hou, C. Cao, X. Liu, and Y. Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pages 382–398. Springer, 2020.

[11] Z. Huang, D. Xue, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua. 3d local convolutional neural networks for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14920–14929, 2021.

[12] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger. Gait-based recognition of humans using continuous hmms. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 336–341. IEEE, 2002.

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

[14] C. P. Lau, C. D. Castillo, and R. Chellappa. Atfacegan: Single face semantic aware image restoration and recognition from atmospheric turbulence. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):240–251, 2021.

[15] C. P. Lau, H. Souri, and R. Chellappa. Atfacegan: Single face image restoration and recognition from atmospheric turbulence. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 32–39. IEEE, 2020.

[16] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3958–3967, 2019.

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

[18] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021.

[19] Y. Liu, R. Collins, and Y. Tsin. Gait sequence analysis using frieze patterns. In *European Conference on Computer Vision*, pages 657–671. Springer, 2002.

[20] Y. Liu, R. T. Collins, and Y. Tsin. A computational model for periodic pattern perception based on frieze and wallpaper groups. *IEEE transactions on pattern analysis and machine intelligence*, 26(3):354–371, 2004.

[21] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006.

[22] S. A. Niyogi and E. H. Adelson. Analyzing gait with spatiotemporal surfaces. In *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 64–69. IEEE, 1994.

[23] Y. Peng, S. Hou, K. Ma, Y. Zhang, Y. Huang, and Z. He. Learning rich features for gait recognition by integrating skeletons and silhouettes. *arXiv preprint arXiv:2110.13408*, 2021.

[24] Y. Ran, Q. Zheng, R. Chellappa, and T. M. Strat. Applications of a simple characterization of human gait in surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):1009–1020, 2010.

[25] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1569–1577, 2022.

[26] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318. IEEE, 2021.

[27] L. Wang and J. Chen. A two-branch neural network for gait recognition. *arXiv preprint arXiv:2202.10645*, 2022.

[28] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[29] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):209–226, 2016.

[30] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2899–2908, 2020.

[31] R. Yasarla and V. M. Patel. Learning to restore a single face image degraded by atmospheric turbulence using cnns. *arXiv preprint arXiv:2007.08404*, 2020.

[32] R. Yasarla and V. M. Patel. Learning to restore images degraded by atmospheric turbulence using uncertainty. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1694–1698. IEEE, 2021.

[33] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006.

[34] G. Zhao, G. Liu, H. Li, and M. Pietikainen. 3d gait recognition using multiple cameras. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 529–534. IEEE, 2006.

[35] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4913–4922, 2019.

[36] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *European Conference on Computer Vision*, pages 681–697. Springer, 2002.

[37] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, 2004.

[38] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4747–4756, 2017.