Person Re-Identification from Gait using an Autocorrelation Network

Cassandra Carley Ergys Ristani Carlo Tomasi
Duke University
Durham, NC, USA
{carley,ristani,tomasi}@cs.duke.edu

Abstract

We propose a new biometric feature based on autocorrelation using an end-to-end trained network to capture human gait from different viewpoints. Our method condenses an unbounded image stream into a fixed size descriptor, and capitalizes on the periodic nature of walking to leverage sequence self-similarity. Autocorrelation is invariant to start or end of the gait cycle, can be efficiently computed online, and is well suited for capturing pose frequencies. We demonstrate empirically that under equal settings an autocorrelation network provides a more complete representation for gait than existing work, resulting in improved person re-identification performance.

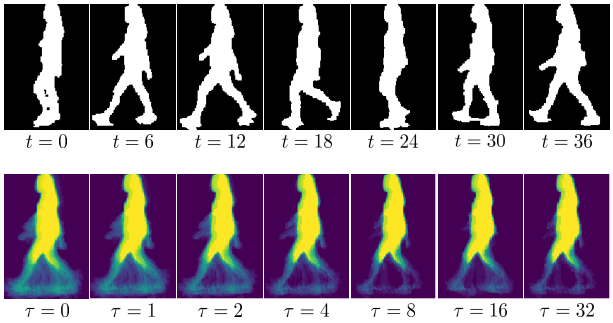


Figure 1: *Top*: Silhouette sequence sampled at different time steps. *Bottom*: Sequence autocorrelation for varying time lags. Autocorrelation at time lag zero is equivalent to the Gait Energy Image, while other time lags capture additional aspects of gait.

1. Introduction

Person re-identification from gait aims to re-identify people who walk based on motion patterns from video, without information about people's visual appearance. Re-identifying people at real-time speeds is of paramount importance for numerous applications, including surveillance, activity understanding, and anomaly detection.

The sequence of images of a walking person captures their gait, and throughout this paper we refer to gait as the cyclic motion that repeats at a stable frequency [5]. The nature of video surveillance makes gait particularly appealing to handle subjects that are seen from a distance. However, the problem is challenging due to changes in viewpoint, appearance (clothing, accessories, and background), walking speed, and duration [9, 17, 28].

Re-identifying people by gait complements appearance-based methods. Appearance-based methods utilize information from one or multiple images but the order is often not relevant. Invariance to order in time makes appearance methods focus on the overall shape of a person and the different poses attained at different times of the walking cycle, without specific knowledge of *when* in the cycle a particular pose is attained. In contrast, gait captures patterns which describe how the shape and pose of a person *change* over time. This class of methods is oblivious to how a person looks but instead it requires a sequential ordering of frames to identify people through changes in appearance induced by their motion. This is particularly useful when people have roughly the same shape and outfit, like sport players, and who can be disambiguated only while walking.

Simple and effective methods like the Gait Energy Image (GEI) have become popular gait descriptors over the last several years, and more recently supervised learning methods have shown improved gait identification accuracy with the availability of larger datasets. Existing methods however resort to a pre-processing step before training and inference, to segment and sample one walking cycle from a walking sequence. Additionally, the motion patterns, whether learned or engineered, do not capture repetitive motion information *explicitly*.

In this work we propose to re-identify people walking based on their gait patterns using the concept of autocorrelation, that is, the similarity of a temporal signal with

This material is based upon work supported by the National Science Foundation under Grant No. 1513816.

its time-shifted self. We treat the intensity of each feature map pixel over time as a signal whose pattern reveals gait traits, and the combination of such signals over the entire image domain allows us to learn a representation of gait across individuals (Figure 1). Computationally, autocorrelation can be implemented efficiently using the Fast Fourier Transform and it can be updated online through recursive estimation. Additionally, our feature can be learned or hand-crafted from any set of raw data observations or intermediate representations. Further, autocorrelation is able to convert walking sequences of variable length into a fixed-size feature, without resorting to pre-processing techniques. In our analysis we also show that the popular GEI descriptor can be viewed as a special case of autocorrelation when the input is a sequence of binary silhouettes, and we demonstrate that under equal settings autocorrelation outperforms existing gait representations.

2. Related Work

Many existing approaches to person re-identification (ReID) by gait are based on two assumptions [5]. The first is that the sequence of configurations for human walking is similar across most people—that is, in general their arms and legs tend to swing forward and backward in similar ways during walking [5]. The second assumption is that, these similarities notwithstanding, there are differences across people, such as those in the length of the arms and legs, the shape of the body, and the period and detailed sequence of the walking cycle configurations [5].

Features for gait recognition can be extracted from each frame and concatenated into a corresponding feature sequence [3, 5, 15, 25]. Alternatively, some methods use a feature extracted from the correlation of a set of frames without considering order [5, 8]. Gait recognition then requires defining a gait feature and a method to compare features, which can be implemented either using statistics about a feature sequence or by comparing frame-level features from a feature sequence that is time-normalized by the walking cycle period [5]. We take a statistics-based approach, using autocorrelation to fix the length of our feature vector in time without requiring the detection of the gait period.

A recent article [1] studied best practices for ReID, based on an experimental comparison of different methods. Largely, approaches are either generative or discriminative. From a pair of co-identical gait features from different viewpoints the generative methods first generate the features for the same viewpoint to enable better matching [11, 18, 28]. However, generative approaches are focused on the accuracy of the generated gait features rather than their discrimination capability [28]. By contrast, discriminative approaches focus on optimizing discrimination capability across viewpoints. This is done by learning a discriminative feature or metric and is typically implemented with machine learning.

The Gait Energy Image (GEI), or averaged silhouette, is the most commonly used gait feature [5, 28]. This feature was thus named for several reasons. First, each frame-level silhouette gives a space-normalized energy image for the person walking at that time-step. Next, the GEI gives the "time-normalized accumulative energy image" for the complete walking cycle of the person. Finally, the GEI can be interpreted such that pixels of higher intensity (energy) correspond to more frequent positions of the person walking.

While the GEI is the most common gait feature, a survey [16] describes other energy-based features for gait. These methods require extracting the gait period and many, including GEI, do so by employing the concept of autocorrelation. Our work avoids the need for detecting the gait period by building on approaches that leverage autocorrelation [12, 27] directly to compute a fixed-size feature from any intermediate representation, even those with varying extents in space and time.

As discriminative methods that use CNNs have become increasingly popular, several gait recognition methods have shown improvement of a CNN-based approach over a baseline without a CNN [26, 28, 31].

Recent progress has been made using deep learning and residual networks [6], and learning hierarchical discriminative features with several levels of granularity [30]. Siamese convolutional neural network (SCNN) approaches use two parallel CNNs that share parameters and a loss considering co-identical and non co-identical pairs of features [32]. In [28] a simple CNN is used with gait energy images (1in-GEINet) to achieve state-of-the-art Rank-1 performance on OU-MVLP, the largest data set for gait recognition .

For methods that do not use CNNs, discriminative features from a GEI extracted by PCA and further reduced by LDA [21] were found to yield slight improvement over directly matching GEIs [28]. Additionally, several generative approaches employ a view transformation model (VTM) [13, 14, 18, 20] to help mitigate view variations from spatial displacements of limbs [28].

3. Method

We model observations $\mathbf{x}(t)$ of a moving person over time as a discrete stochastic process, that is, a mapping from an event space $\mathcal{E}$ to a space $\mathcal{X}$ of real-valued vector functions defined on the integers. Given an event $e \in \mathcal{E}$, a realization of the stochastic process is thus a function:

$$\mathbf{x}_e(t) \; : \; \mathbb{Z} \rightarrow \mathbb{R}^N \tag{1}$$

and the subscript denoting dependence on e is typically omitted for simplicity. We think of the integer variable t as denoting *time*, and the vector $\mathbf{x}(t)$ for a given t describes frame t in some fashion.

The process $\mathbf{x}(t)$ is said to be *stationary* if this density depends only on the differences $\tau_j \stackrel{\text{def}}{=} t_j - t_0$ for $j = 1, \ldots, k$ and not on $t_0, \ldots, t_k$.

The process is said to be *Wide-Sense Stationary* (WSS) if the analogous property holds for the mean and the autocorrelation, that is for the first and second moments for $k = 1$ and $k = 2$. Some human activities are at least approximately WSS, and are then described in the literature as *motion textures*. In the short term, the quasi-periodic nature of human gait makes walking a *cyclostationary* process, that is, one whose statistics are periodic. However, if walking is observed over several time intervals starting at random times during the gait cycle, the cyclic nature of the activity averages out over different observations, and yields an approximately stationary process.

Our proposal is to use the autocorrelation $A(\tau)$ for a set of time lags τ as a descriptor of human gait, where autocorrelation is defined as:

$$A(\tau) = \mathbb{E}[\mathbf{x}(t)\mathbf{x}^T(t+\tau)] \; . \tag{2}$$

If the input process $\mathbf{x}(t)$ is N-dimensional and L time lags are considered, the values of the autocorrelation can be stored in a block of data of size $N \times N \times L$. When N is very large, we may only use the diagonal of $A(\tau)$.

Time Lags. The time lags under consideration will be included in a single vector:

$$\boldsymbol{\tau} \stackrel{\text{def}}{=} [\tau_1, \ldots, \tau_L] \in \mathbb{Z}^L \; . \tag{3}$$

The values of these time lags are between 0 and some maximum lag, $\tau^{max} = \max \boldsymbol{\tau}$, and they are an ordered subset of $\{0, \ldots, \tau^{max}\}$. An interesting option is to compute the autocorrelation for all low lag-values and then sample higher values with logarithmic density, that is, with exponentially growing intervals between lag values [2, 4].

Properties. It is easy to verify that the matrix $A(0)$, that is, the second moment of $\mathbf{x}(t)$ for any t, is semidefinite positive, and that for any $\tau \in \mathbb{Z}$ the following equality holds:

$$A(\tau) = A^T(-\tau) \; . \tag{4}$$

The same properties hold for the autocovariance $\Sigma(\tau)$. Positive-semidefiniteness reflects the fact that $A(0)$ is a power and $\Sigma(0)$ is a covariance. Since $A(0)$ is the correlation of the signal with itself, we also have $\|A(\tau)\|_2 \leq \|A(0)\|_2$ for all τ. Further properties of autocorrelation or autocovariance can be found in standard texts [4, 22, 23].

Recursive Computation. The autocorrelation $A(\tau)$ is an expectation. If the process $\mathbf{x}(t)$ is WSS and ergodic, the expectation of any quantity $q(t)$ related to $\mathbf{x}(t)$ can be estimated from the empirical average of a sufficiently large number of samples:

$$\mathbb{E}[q(t)] \approx \mu(t) \stackrel{\text{def}}{=} \frac{1}{|T(t)|} \sum_{t \in T(t)} q(t) \tag{5}$$

where:

$$T(t) \stackrel{\text{def}}{=} \{t_1, \ldots, t_{|T(t)|}\} = \{s \in \mathbb{Z} \mid s \leq t\} \tag{6}$$

is the set of $|T(t)|$ time tics elapsed up until time t. For simplicity, T will be used to denote the number of time samples used, so that $T = |T(t)|$. We consider scalar quantities here: For vectors or matrices, the discussion is merely repeated for each component.

For video sequences with unbounded time extents, one needs a way to update $\mu(t)$ based on the information available up to time t. Because the data varies, μ depends on t even if the underlying process is stationary: While the true mean m may be constant, its estimate $\mu(t)$ varies. For efficiency, $\mu(t)$ can be computed as a (weighted) moving average, for which a recursion with finite state can be defined.

Implementation. Given a sequence of N images or feature maps each with C channels of size $W \times H$, we compute the autocorrelation $A(\tau)$ for each pixel and time lag. The result is a matrix of size $W \times H \times \tau^{max}$ that does not depend on the (varying) number of images N. We also integrate autocorrelation as an operator into a CNN, and the operator is differentiable and has no learnable parameters. For binary silhouettes, $A(0)$ is simply the GEI of the sequence.

4. Experiments

We examine empirically what we can tell about a person's identity from walking patterns. Specifically, the input is a silhouette video clip of a person who walks across the field of view of a camera placed at an unknown viewpoint. The output is a ranking of gallery observations and their corresponding identities, in order of similarity to the input observation.

We use human motion here in a very indirect manner for inference: The individuals included in the training set are different from those in the test set, so the problem is not to recognize a person from his or her gait (gait recognition), but rather for the system to understand what aspects of someone's gait are most helpful in recognizing identity (re-identification from gait).

We use the OU-MVLP data set [28] for evaluation based on input of a sequence of silhouettes and show results that improve on the prior state-of-the art, comparing variants of our proposed autocorrelation motion texture feature with the popular Gait Energy Image (GEI) feature [5] and 3D convolutional neural networks (C3D) [29].

Next we describe the details for running our experiments. First, we detail the setup, benchmark data set and performance measures used for evaluation. Then we discuss the methods we compare to each other and implementation details for training and testing.

4.1. Setup

Given an image of a person (the *query*), a person re-identification system retrieves from a database a list of other images of people (*gallery*), usually taken from different cameras and at different times, and ranks them by decreasing similarity to the query. Ideally, any images in the database that are *co-identical* with (that is, depict the same person as) the person in the query are ranked highly. Rather than using a single image, we use a sequence of size-normalized silhouettes to compute our autocorrelation motion texture feature to perform person re-identification from gait. The intention is that features extracted from the same person are more similar to each other than to those extracted from different people. The sets of people observed for training, validation, and testing are mutually disjoint.

4.1.1 Benchmarks

We perform experiments on the OU-ISIR multi-view large population data set (OU-MVLP), which is the largest multi-view gait database available to date and can be used to evaluate person re-identification from gait [28]. The data set has $10,307$ subjects with varying ages (2-87 years old) and genders. Video is captured from 14 view angles ($0° - 90°, 180° - 270°$, Fig. 2a) for two separate walking sequences (A run from A to B and B run from B to A), giving 28 gait image sequences per subject. Including a second sequence for each subject implies that the data set includes view variations and intra-subject variations of gait itself [28]. The original sequences for OU-MVLP are composed of binary images of 1280×980 pixels and captured at 25 fps. The binary images show a silhouette region that was extracted using a chroma-key method to remove the green background from the controlled walking course (Fig. 2b).

4.1.2 Evaluation

For each identity in the test set, we use the specified model to compute its corresponding K_f-dimensional feature for each of its 28 samples (run A and run B at each of the 14 viewpoints). To measure performance, all A runs are used as the query set and B runs as the gallery set. Given query viewpoint(s) and gallery viewpoint(s), the feature distances are computed and sorted to give the identities of the gallery set in decreasing similarity for each query. We report Average Rank and Rank-1 accuracy per viewpoint. We also report these scores averaged across all queries and viewpoints. Average Rank gives the mean position of the correct identity. Rank-N accuracy is the percentage of queries that return the correct gallery identity within the top N results. For Rank-1, the probability of a random guess being accurate is 1 in 1800 (the number of gallery identities). Thus, the random baseline accuracy per viewpoint pair is 0.056%.

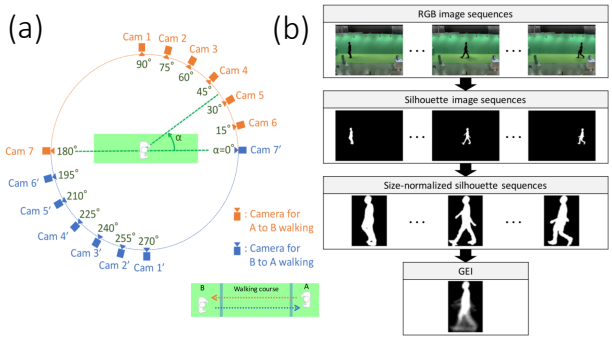


Figure 2: Camera setup for OU-MVLP gait data set [28] (a) and their method for extracting same-sized silhouette sequences and computing the GEI (b).

4.1.3 Preprocessing

We follow the method by the authors of OU-MVLP to extract size-normalized 128×88 silhouettes for each binary image (Fig. 2, [28]). Each frame has a corresponding silhouette region, and the normalization step first extracts its top, bottom, and horizontal center. The top and bottom are defined as the extremum of the sorted y coordinates of the silhouette region, while the horizontal center is set as the median of the sorted x coordinates of the region. A moving-average filter is then applied to these positions [10]. A 128×88 silhouette image is then produced such that the horizontal median of the silhouette region corresponds to the horizontal center of the silhouette image.

4.1.4 Compared Methods

Given the size-normalized silhouette sequences we describe several models for comparing our gait re-identification method ACnet with other state-of-the-art models.

GEI. We compare with the implementation of [28] (Fig. 3(a)) which is a CNN-based method using a single Gait Energy Image [5] as input (1in-GEInet). This approach has achieved state-of-the-art Rank-1 performance on the OU-MVLP dataset. We follow the approach used by the OU-MVLP authors to extract a Gait Energy Image (GEI) (Fig. 2(b)), or the pixel-wise average of the size-normalized silhouette sequences over one gait period [28]. The *gait period* is detected using the normalized autocorrelation (NAC) of the size-normalized silhouette sequences from the side-

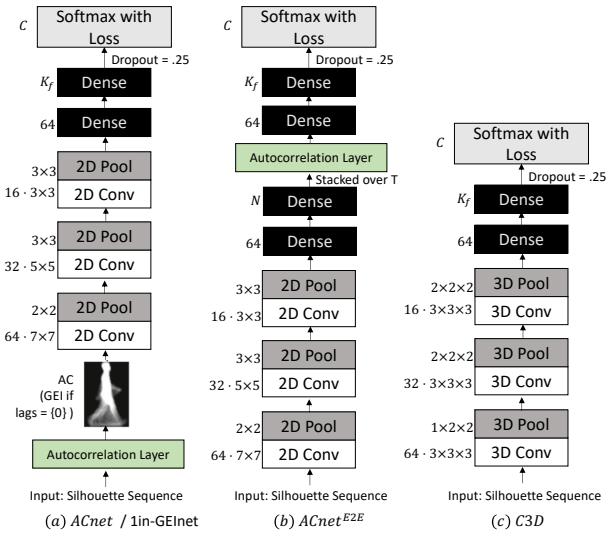


Figure 3: Architectures for the different methods we compare for Re-ID by gait. Numbers written to left side of Conv, Pool, and Dense indicate: [#filters · filter size], [filter size], and [# of nodes] respectively. K_f is the final embedding dimension of the feature at the last layer of the network before an optional softmax is used to produce a C-dimensional vector, where C is the number of subjects.

view camera (90°) [10]. Specifically, the gait period is identified as time shift corresponding to the second peak of the NAC [10]. If multiple gait periods are detected, [28] use the period closest to the center of the walking course. We use the gait period to directly compare with 1in-GEInet, but perform most of our experiments on sequences that start at any point in the gait cycle and use different temporal windows that are not the exact gait period.

C3D. We also compare against the popular Convolution 3D (C3D) CNN architecture, which has previously achieved state-of-the-art results for both action recognition and person re-identification by gait with silhouettes [29]. Figure 3(c) shows our C3D implementation, which matches standard state-of-the-art C3D architectures, and the results we report are consistent with [29].

ACnet. We compare two models of our `ACnet`. The standard version uses the autocorrelation layer on the input of same-sized silhouettes to compute the autocorrelation feature which is then fed to a CNN, with architecture shown in Figure 3(a). The standard version of `ACnet` with $\tau = 0$ (`ACnet`$\{0\}$) is our implementation of 1in-GEInet [28]. We also implement a version of our autocorrelation network, `ACnet`$^{\text{E2E}}$, that is trained end-to-end (E2E) from image sequences as shown in Figure 3(b). The learned `ACnet` (`ACnet`$^{\text{E2E}}$) model uses an additional shallow network before the autocorrelation layer and is trained end-to-end to learn a more abstract frame-level feature.

4.1.5 Training and Testing

The networks are implemented using Keras. A training and validation set are specified along with the number of epochs and batch size to use for training. Unless otherwise specified, we use the following default parameters and settings for the inputs and network structures in our experiments.

For training, testing, and validation we use disjoint sets of $1800, 1800$, and 400 identities respectively. The initial inputs are a sequence of $T = 100$ size-normalized 128×88 silhouettes. The final embedding dimension of the feature at the last layer of the network (before an optional softmax) is $K_f = 32$. When the optional softmax is used, the output is a C-dimensional vector, where C is the number of classes, in this case the number of identities.

We follow the approach introduced by [7] of using PK batches for training. Each batch consists of P identities with K samples per identity selected at random. At each training epoch, every identity is selected and a batch constructed by choosing the other $P - 1$ identities at random. This approach avoids the need to generate a combinatorial number of triplets, and is well suited to similarity-based ranking tasks [24].

For training, we set the batch size to 50 ($P = 10$ identities, $K = 5$ samples). The query and gallery viewpoints are selected at random from all 14 viewpoints, similar to prior work [28]. The learning rate is set to $3 \cdot 10^{-4}$ for the first 10 epochs, $3 \cdot 10^{-5}$ for the next five epochs, and then $3 \cdot 10^{-5}$ thereafter.

We consider two types of losses, categorical cross-entropy and hard triplet loss with adaptive weighting [24]. The hard triplet loss requires that for each anchor identity, the furthest sample from the same identity has smaller distance than the nearest sample from other identities. Additionally, we use a multi-task loss (denoted by +m) that combines the two.

We train the models until convergence, using an early stopping patience of 5 epochs (with validation loss) and a maximum number of epochs of 100. For multi-loss we do not implement validation and thus stop at a specified epoch (100). For all experiments we report results from the best epoch (using minimum validation loss).

4.2. Results

We present results for the task of person re-identification from gait using the OU-MVLP data set. We perform several experiments and discuss results using different measures (Rank-1, Avg. Rank), highlighting where our method outperforms prior state-of-the-art; examine the influence of

Method	Rank-1 (%) ↑	Avg. Rank ↓
C3D [29]	21.3	41
C3D+m [29]	30.0	31
1in-GEInet [28]	36.1	19
1in-GEInet+m [28]	36.9	24
$\texttt{ACnet}^{\text{E2E}}\{7log\}$+m	**58.0**	**7**

Table 1: State-of-the-art results for Re-ID by gait, showing the improvement of our autocorrelation network (ACnet) using a set of $\{7log\}$ time lags and trained end-to-end (E2E) with multi-loss (+m) over prior state of the art.

different pipeline components; and finally analyze system limitations.

4.2.1 Comparisons with the State of the Art

Table 1 shows the average Rank-1 performance of state-of-the-art Re-ID by gait methods on OU-MVLP across 14 viewpoints ($0° - 270°$) for $1,800$ identities. Our method $\texttt{ACnet}^{\text{E2E}}\{7log\}$+m achieves 58.0% Rank-1 performance. This result is significant considering the random baseline (0.056%). Further, we significantly outperform the prior state of the art. Our Rank-1 performance (58.0%) shows significant improvement over the popular C3D architecture, trained with single-loss (21.3%) and multi-loss (30.0%) (Table 1, and Fig. 5 over Fig. 4(c)). Due to its architecture, C3D can only handle small batches of size 20 ($P = 5, K = 4$) and temporal samples $T = 15$. While this batch size is limiting, we allow C3D to run until convergence. Therefore, the small temporal window is more likely the factor limiting C3D's performance.

Our experiments show improvement of our $\texttt{ACnet}^{\text{E2E}}\{7log\}$+m method ($58.0\%$) over 1in-GEInet, equivalent to our $\texttt{ACnet}\{0\}$, trained with single-loss (36.1%) and multi-loss (36.9%) (Table 1). Note that [28] report mean Rank-1 accuracy of 40.7%, but that this is for query and gallery angle pairs from a thinned set of viewpoints: $[0°, 30°, 60°, 90°]$. We use this thinned set of viewpoints to provide further detailed results for our method and those compared (See Fig. 4 and 5), and report a similar performance for our implementation of 1in-GEInet over the thinned viewpoints (40.6%) (Fig. 4(a)).

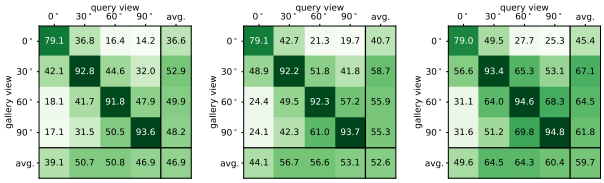
While we only report results for Rank-1, our performance is comparable or surpasses state-of-the-art methods that relax the challenge by using Rank-5, looking only at intra-view test sets, or limiting the viewpoints to $55° - 90°$.

4.2.2 Model Parameters and Architecture

We further compare our method with 1in-GEInet and explain their similarities while detailing the modifications of our approach that lead to improved performance. Using

(a) $\texttt{ACnet}\{0\}$ (1in-GEInet) (b) $\texttt{ACnet}\{0\}$+m (c) C3D+m

Figure 4: Rank-1 results for prior state of the art, for (a) our implementation of [28]'s 1in-GEInet with single loss, (b) multi-loss, and (c) [29]'s C3D with multi-loss.



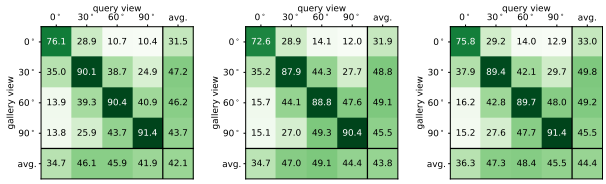
(a) $\texttt{ACnet}^{\text{E2E}}\{0\}$ (b) $\texttt{ACnet}^{\text{E2E}}\{7log\}$ (c) $\texttt{ACnet}^{\text{E2E}}\{7log\}$ +m

Figure 5: Rank-1 results showing our improvements of using ACnet with end-to-end (E2E) training, $\{7log\}$ time lags, and multi-loss (+m).

Method	Rank-1 (%) ↑	Avg. Rank ↓
$\texttt{ACnet}\{0\}$	36.1	19
$\texttt{ACnet}\{0, 32\}$	36.9	18
$\texttt{ACnet}\{7step\}$	39.4	**16**
$\texttt{ACnet}\{7log\}$	**39.6**	**16**

Table 2: Re-ID by gait results on OU-MVLP, showing improvement of different sets of time lags.

a single time lag of zero is equivalent to directly comparing the signals with themselves without a time lag. In the case of ACnet the input to the autocorrelation function is the size-normalized silhouettes. Each pixel is treated as a separate signal, with observations of one or zero over the set of time samples. The autocorrelation of a single pixel signal is then the summation of itself multiplied by a version of itself that is circularly shifted by the given time lag, normalized by the number of time samples. For a time lag of zero the autocorrelation is simply the summation of the observations normalized by the number of time samples. Thus, the autocorrelation at $\tau = \{0\}$ of the direct silhouette input is equivalent to the GEI as long as the observations are from the set of time samples that correspond to a single complete gait period. Therefore, our $\texttt{ACnet}\{0\}$ is an implementation of 1in-GEInet [28].



(a) ACnet$\{0, 32\}$ (b) ACnet$\{7step\}$ (c) ACnet$\{7log\}$

Figure 6: Rank-1 results for ACnet with different time lags.

Method	Rank-1 (%) ↑	Avg. Rank ↓
ACnet$\{0\}$	36.1	19
ACnet$^{\text{E2E}}\{0\}$	**41.6**	**11**
ACnet$\{7log\}$	39.4	16
ACnet$^{\text{E2E}}\{7log\}$	**49.7**	**8**
ACnet$\{7log\}$+m	56.1	10
ACnet$^{\text{E2E}}\{7log\}$+m	**58.0**	**7**

Table 3: Re-ID by gait results on OU-MVLP, showing improvement of using end-to-end (E2E) training for ACnet with time lags of $\{0\}$ and $\{7log\}$, with single loss and multi-loss (+m).

Different Time Lags. Using multiple time lags improves the average Rank-1 performance as shown in Table 2 and Figure 6. In Table 2, we study the average Rank-1 performance over all 14 viewpoints for different sets of time lags used with the standard ACnet. The baseline performance with a single $\tau = \{0\}$ (36.1%) is improved only slightly (36.9%) by considering the pair of lags $\{0, 32\}$. This makes sense, as the typical number of time samples in a single complete gait period was approximately 32 time-steps, and using a time lag that is the exact length of the gait period is equivalent to a time lag of zero. Additionally, we considered two sets of time lags, $\{7log\} = \{0, 1, 2, 4, 8, 16, 32\}$ and $\{7step\} = \{0, 10, 20, 30, 40, 50, 60\}$. A significant improvement over the baseline is achieved by both $\{7log\}$ (39.6%) and $\{7step\}$ (39.4%). The similar performance between $\{7log\}$ and $\{7step\}$ is again likely due to the gait period length ≈ 32 meaning that lags $40, 50$, and 60 do not capture significantly different values than those in $\{0, 32\}$.

We can see that adding multiple time lags seems to help mostly the off-diagonal performance values for cross-view tasks (Fig. 6(b and c over a) and over Fig. 4(a)). However, this might simply be because the intra-view performance has been saturated, given that it is a much easier problem.

End-to-End Training. Placing the autocorrelation layer in the middle of the CNN and training end-to-end (E2E) significantly improves performance for ACnet$\{0\}$ (36.1% to 41.6%), ACnet$\{7log\}$ (39.4% to 49.7%), and ACnet$\{7log\}$+m (56.1% to 58.0%) as shown in Table 3.

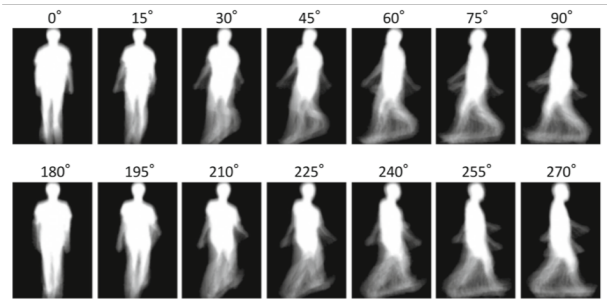


Figure 7: Autocorrelation for time lag 0, equivalent to GEI [5], shown for different viewpoints.

Method	Rank-1 (%) ↑	Avg. Rank ↓
ACnet$\{0\}$	36.1	**19**
ACnet$\{0\}$+m	**36.9**	24
ACnet$\{7log\}$	39.4	16
ACnet$\{7log\}$+m	**56.1**	**10**
ACnet$^{\text{E2E}}\{7log\}$	49.7	8
ACnet$^{\text{E2E}}\{7log\}$+m	**58.0**	**7**
C3D [29]	21.3	41
C3D+m	**30.0**	**31**

Table 4: Re-ID by gait results on OU-MVLP show improvements from training with multi-loss (+m is triplet + categorical) over single-loss (triplet) for C3D and our ACnet with time lags $\{0\}$ and $\{7log\}$ and trained end-to-end (E2E).

A Direct Matching (DM) approach that simply compares feature distance on the GEI directly (flattening it into a 1D vector and using L2 distance) was found to only work for same-viewpoint pairs [28]. This is because the GEI captures the spatial displacement of the limbs and is a direct function of the viewpoint, such that those closer to side-view demonstrate maximum front-to-back displacement while those closer to front-view display maximum side-to-side displacement (Fig. 7). Hence, the GEI has large intra-sample variation across viewpoints. This suggests that adding a shallow CNN to the GEI can help learn a view-invariant feature and supports the improvement we see in our end-to-end approach (Table 3). Our results would be bolstered by future experiments that control for consistency in the number of dense layers across methods as well as testing other hand-crafted features as input to the end-to-end network.

Different Losses. Training models with a multi-loss (+m is triplet + categorical) improves performance over using a single loss (triplet), as shown in Table 4 for several architectures: ACnet$\{0\}$ (36.1% to 36.9%), ACnet$\{7log\}$ (39.4% to 49.7%), ACnet$^{\text{E2E}}$ $\{7log\}$ (56.1% to 58.0%), and [29]'s C3D (21.3% to 30.0%).

Temporal Window	Rank-1 (%) ↑	Avg. Rank ↓
$T = 40$	49.3	15
$T = 100$	56.1	**10**
$T = 150$	**56.3**	**10**

Table 5: Re-ID by gait results on OU-MVLP show improvements from using a longer temporal window ($T = 100$ v. $T = 40$) but with marginal returns extending further ($T = 150$) for ACnet$\{7log\}$+m, our ACnet with $\{7log\}$ time lags and multi-loss (+m).

Temporal Window. Increasing the temporal window improves performance, as shown in Table 5 and Figure 8. Using our ACnet$\{7log\}$+m we show that increasing the temporal window from $T = 40$ to 100 improves Rank-1 performance across all 14 viewpoints (from 49.3% to 56.1%). A further increase to $T = 150$ only slightly improves performance (56.3%). This shows that there is a point beyond which extending the temporal window is no longer beneficial, and thus for the majority of our experiments we consider $T = 100$.

Given that the average gait period is $T = 32$, these results also suggest that for periodic activities such as walking it is useful to capture multiple cycles of the activity. Comparing these results along with the performance of C3D, which can only handle 15 time samples for the same memory budget, autocorrelation proves to be an efficient and compact way to consider a larger temporal window.

(a) $T = 40$ (b) $T = 100$ (c) $T = 150$

Figure 8: Rank-1 results for different temporal window lengths (T) for ACnet$\{7log\}$+m, our ACnet with $\{7log\}$ time lags and multi-loss (+m).

Intra-View v. Cross-View. The diagonal of the Rank-1 matrices (Fig. 9) represents the less difficult *intra-view* task of learning similarity within the same camera viewpoint. By contrast, the off-diagonal represents *inter-view* pairs and entails the much more challenging task of cross-view learning. It seems that the query-gallery pair $(90°, 0°)$ is one of the most challenging inter-view pairs, which is intuitive considering that front- and side-views are the most separate viewpoints in terms of appearance from a single frame and from the sequence of frames. Further, we notice that it is particularly challenging in general when one of the viewpoints is

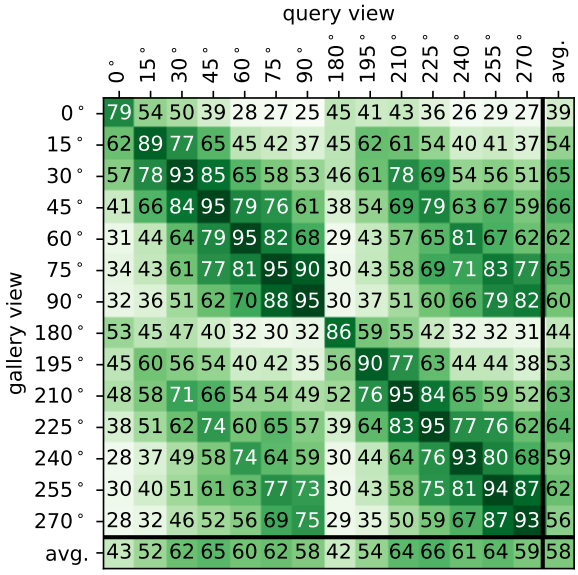


Figure 9: State-of-the-art Re-ID by gait Rank-1 results for ACnet$^{E2E}\{7log\}$+m, our model using $\{7log\}$ time lags and trained end-to-end (E2E) with multi-loss (+m) on OU-MVLP for all viewpoints.

straight on ($0°$ or $180°$) as there is less spatial displacement from the limbs and hence less signal to re-identify from gait.

The Rank-1 matrices for all viewpoints seem as though they can be divided into four sub-matrices that have similar trends, as seen in Figure 9. We notice, as did [28], a similarity between (query, gallery) pairs separated by $180°$. The implementation flips one of the viewpoints so they are similar to a "same-view pair due to perspective projection assumption" [19]. This explains why the trends are similar in the sub-matrices defined by: (a) $0° - 90°$ query to $0° - 90°$ gallery; (b) $180° - 270°$ query to $0° - 90°$ gallery; (c) $0° - 90°$ query to $180° - 270°$ gallery; and (d) $180° - 270°$ query to $180° - 270°$ gallery viewpoints.

5. Conclusion

We have introduced a new biometric feature to capture gait autocorrelation leveraging end-to-end training. This feature, whether applied to raw data or intermediate learned representations, captures the time-varying aspects of human gait explicitly and can be used to complement appearance-based re-identification methods. In our experiments we have shown how different parameters affect our method, and how autocorrelation and end-to-end training improves performance over existing methods under equal settings. In future work, we plan to apply autocorrelation to 2D and 3D pose sequences, as well as raw RGB sequences of people walking.

References

[1] J. Almazan, B. Gajic, N. Murray, and D. Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339*, 2018. 2

[2] P. H. Colberg and F. Höfling. Highly accelerated simulations of glassy dynamics using gpus: Caveats on limited floating-point precision. *Computer Physics Communications*, 182(5):1120–1129, 2011. 3

[3] R. T. Collins. Multitarget data association with higher-order motion models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1744–1751. IEEE, 2012. 2

[4] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001. 3

[5] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2006. 1, 2, 3, 4, 7

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[8] P. S. Huang, C. J. Harris, and M. S. Nixon. Recognising humans by gait via parametric canonical space. *Artificial Intelligence in Engineering*, 13(4):359–366, 1999. 2

[9] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi. Gait verification system for criminal investigation. *Information and Media Technologies*, 8(4):1187–1199, 2013. 1

[10] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5):1511–1521, 2012. 4, 5

[11] A. Kale, A. R. Chowdhury, and R. Chellappa. Towards a view invariant gait recognition algorithm. In *Advanced Video and Signal Based Surveillance, 2003. Proceedings. IEEE Conference on*, pages 143–150. IEEE, 2003. 2

[12] T. Kobayashi and N. Otsu. A three-way autocorrelation based approach to human identification by gait. 2006. 2

[13] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1058–1064. IEEE, 2009. 2

[14] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Gait recognition under various viewing angles based on correlated motion regression. *IEEE transactions on circuits and systems for video technology*, 22(6):966–980, 2012. 2

[15] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, 1(2):1–32, 1998. 2

[16] Z. Lv, X. Xing, K. Wang, and D. Guan. Class energy image analysis for video sensor-based gait recognition: A review. *Sensors*, 15(1):932–964, 2015. 2

[17] N. Lynnerup and P. K. Larsen. Gait as evidence. *IET biometrics*, 3(2):47–54, 2014. 1

[18] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *European Conference on Computer Vision*, pages 151–163. Springer, 2006. 2

[19] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Which reference view is effective for gait identification using a view transformation model? In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 45–45. IEEE, 2006. 8

[20] D. Muramatsu, Y. Makihara, and Y. Yagi. Cross-view gait recognition by fusion of multiple transformation consistency measures. *IET Biometrics*, 4(2):62–73, 2015. 2

[21] N. Otsu. Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In *Proceedings of the 6th International Conference on Pattern Recognition, 1982*, pages 557–560, 1982. 2

[22] A. Papoulis and S. U. Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002. 3

[23] J. G. Proakis. *Digital signal processing: principles algorithms and applications*. Pearson Education India, 2001. 3

[24] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. *arXiv preprint arXiv:1803.10859*, 2018. 5

[25] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):162–177, 2005. 2

[26] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *Biometrics (ICB), 2016 International Conference on*, pages 1–8. IEEE, 2016. 2

[27] D. Skog. Gait-based reidentification of people in urban surveillance video, 2010. 2

[28] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015. 3, 5, 6, 7

[30] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. *arXiv preprint arXiv:1804.01438*, 2018. 2

[31] 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, 2017. 2

[32] C. Zhang, W. Liu, H. Ma, and H. Fu. Siamese neural network based gait recognition for human identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2832–2836. IEEE, 2016. 2