engineering proceedings

MDPI

Proceeding Paper

Enhanced Gait Recognition for Person Identification Using Spatio-Temporal Features and an Attention-Based Deep Learning Model †

Kollaparampil Thomas Thomas [1,2,*] and Kaimadathil Pushpangadan Pushpalatha [1]

[1] School of Computer Sciences, Mahatma Gandhi University, Kottayam 686560, India; pushpalathakp@mgu.ac.in
[2] Department of Computer Science, CHRIST University, Bengaluru 560029, India
* Correspondence: thomas.kt@christuniversity.in
† Presented at the 12th International Electronic Conference on Sensors and Applications, 12–14 November 2025; Available online: https://sciforum.net/event/ECSA-12.

Abstract

Human gait has proven to be one of the standard biometrics for human identification. It is a non-invasive biometric method that uses human walking patterns specific to each human being. In most of the traditional methods, we use handcrafted features of simple convolutional models for gait analysis in human identification. Here, we may face challenges addressing complex temporal dependencies in gait sequences. This study proposes a novel deep learning framework that applies multi-feature input representations. It combines Gait Energy Images (GEIs), Frame Difference Gait Images (FDGIs), and Histogram of Oriented Gradients (HOG) features. This is proposed for enhancing the accuracy of human identification. The proposed work implements a CNN-based feature extractor with an attention mechanism for gait recognition. The model is trained and validated on a labeled dataset, showcasing its ability to learn discriminative gait representations with improved accuracy. The proposed pipeline of activities includes preprocessing and converting gait sequences into frames, organizing them using folder-based numerical extraction, followed by the training of an attention-enhanced convolutional network. The proposed model was found to perform better than existing methods on public datasets and works well even with different camera angles and clothing styles.

Keywords: gait; silhouette; attention mechanism; Gait Energy Image (GEI); Histogram of Oriented Gradients (HOG); Frame Difference Gait Image (FDGI)

check for updates

Academic Editor: Stefano Mariani

Published: 7 November 2025

Citation: Thomas, K.T.; Pushpalatha, K.P. Enhanced Gait Recognition for Person Identification Using Spatio-Temporal Features and an Attention-Based Deep Learning Model. *Eng. Proc.* **2025**, *118*, 101. https://doi.org/10.3390/ECSA-12-26532

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Gait recognition is a promising biometric technology area for human identification. Traditional biometric systems require cooperation from the subject's end, which is not needed in gait biometrics, making it a good choice for surveillance and security applications [1,2]. Gait is the unique way in which a person walks. Most traditional methods like fingerprints, iris scans, or face recognition need active cooperation of the subject. A person's way of walking is unique and can be recorded without their cooperation. This makes gait recognition a powerful method.

There are a lot of applications of gait recognition in today's world, which are vast and growing rapidly:

Surveillance and Public Safety: Gait recognition is used in places like airports, metro stations, and shopping malls to continuously monitor crowds. It helps identify suspects in real-time, even if lighting is poor or the person's face is not clearly visible [1,2].

Forensics and Law Enforcement: Police and investigators use gait analysis to support criminal cases by identifying individuals seen walking in security camera footage [2].

Healthcare and Rehabilitation: In healthcare, analyzing how someone walks helps doctors identify issues related to nerves, muscles, or bones [3].

Workplace and Access Control: Companies use gait recognition for secure, hands-free access to sensitive areas like research labs, military bases, or data centers [3].

Gait analysis has many challenges. Inter-subject variability in the shape of the body, style of the limb movement, or walking pace, etc., may increase complexity. Other factors like clothing, footwear, object carrying conditions, walking surface conditions, etc., may also result in considerable changes in gait patterns [2,4].

This paper studies several types of gait features. The most common, and the ones which are widely used, are the Gait Energy Image (GEI) [5], Frame Difference Gait Image (FDGI), and Histogram of Oriented Gradients (HOG)-based gait representations [6]. As each feature type provides valuable cues, a proper combination of multiple representations can improve the performance by capturing complementary information.

The contributions of this paper are the following:

A unified deep learning architecture is proposed, which integrates three gait representations (GEI, FDGI, and HOG) into a single sequential input pipeline. This has combined both the spatial and temporal features of the gait sequences of a person in 10 different angles.

- A spatial attention mechanism is used to enhance gait recognition performance.
- The effectiveness of the proposed approach is evaluated on a CASIA-A gait dataset, demonstrating improved accuracy compared to conventional CNN baselines.

The structure of the proposed paper is as follows: Section 2 gives the review's related work, followed by Section 3, which gives the details of the proposed methodology. Section 4 shows the experiments and results; Section 5 is the conclusion of the paper, and this section provides the future scope of the work as well.

2. Literature Review

Being non-invasive and mostly unaffected by the distance factor, human recognition through gait analysis has become a topic of increasing research interest. Traditional approaches used in gait analysis are broadly grouped into model-based methods and appearance-based methods. Silhouette representations, like the Gait Energy Image (GEI), have been shown to work really well, amongst the appearance-based procedures [2,5].

GEI, introduced by Han and Bhanu [5], is a spatio-temporal template formed by the method of averaging binary silhouette images over a complete gait cycle. It can be both spatial structure and temporal dynamics in a compact form, making it one of the most widely used gait descriptors. This makes GEI a popular feature utilized for recognizing how people walk. However, GEI does not always work well in changing conditions, like when someone wears different clothes or carries something.

To extract the frequency characteristics of gait patterns, researchers have focused on the Frame Difference Gait Image (FDGI). FDGI obtains periodic motion information by transforming silhouette sequences into the temporal domain. This emphasizes the dominant gait rhythms, which are less affected by visual noise or silhouette distortion [7,8].

Histogram of Oriented Gradients (HOG) is another powerful visual descriptor. HOG was originally developed for object detection [8,9]. HOG captures edge and gradient orientation information, which is needed for detecting body shape and posture. In gait analysis, HOG is obtained from silhouette frames to improve spatial details.

Recent studies using GEI and HOG together have shown promise in capturing both motion energy and spatial texture [10]. However, most of the time, these features are processed independently, which creates a limitation regarding the synergy between spatial and temporal components.

To reduce these issues, researchers have explored deep learning frameworks capable of learning fused representations. It is noted that CNNs have been used extensively to extract high-level spatial features, while Transformer-based architectures have recently shown great ability in capturing long-range temporal dependencies in sequential data [4].

This study proposes a novel deep learning pipeline that combines GEI, FDGI, and HOG features into a single time-sequenced input. A Convolutional Neural Network (CNN) is employed to extract frame-wise spatial features. To improve the performance, an attention mechanism is used, which models the temporal dependencies through the sequence. The design of this integrated framework leverages the complementary capabilities of GEI, FDGI, and HOG, leading to a more robust and accurate gait recognition system under real-life conditions.

3. Methods

3.1. Overview

The methodology proposed in this paper makes use of a robust gait recognition system that integrates three strongly complementary visual features—Gait Energy Image (GEI), Frame Difference Gait Image (FDGI), and Histogram of Oriented Gradients (HOG)—into a spatio-temporal system. This integrated approach is designed in order to obtain rich spatial and temporal dynamics in human gait patterns. The features thus obtained are processed in a hybrid deep learning framework which involves a Convolutional Neural Network (CNN) with a spatial attention mechanism to enhance gait recognition from the image features (128 $\times$ 128 pixels). This architecture improves the discriminative ability and temporal coherence of gait features, thereby improving recognition accuracy in varied conditions.

Model Architecture:

The model begins with an input layer for single-channel images, followed by two convolutional layers (with 32 and 64 filters) and max pooling layers, which extract spatial features and reduce dimensionality. The spatial attention module is then applied, which computes both average and max pooling across the channel dimension, concatenates them, and passes the result through a 7 $\times$ 7 convolution with sigmoid activation, thus producing an attention map. This map is element-wise multiplied with the input highlighting important spatial regions. The attended feature map is further processed by a third convolutional layer with 128 filters, followed by max pooling. After flattening, the features pass through two fully connected layers with dropout regularization to reduce overfitting. At the end, a softMax output layer categorizes the input into one of the gait categories based on the person. The model is compiled using the Adam optimizer and categorical cross-entropy loss for multi-class classification.

3.2. Feature Selection and Its Contributions

To construct a robust and discriminative gait representation, three complementary feature extraction techniques were employed, Gait Energy Image (GEI), Frame Difference Gait Image, and HOG, which support in offering a comprehensive encoding of gait characteristics.

3.2.1. Gait Energy Image (GEI)

GEI is a silhouette-based descriptor that is obtained by calculating the averages of binary gait silhouettes over one complete gait cycle. The binary image, thus obtained, can encode spatial appearance as well as temporal motion patterns. GEI provides more

robustness against intra-class variations and is found to be computationally efficient [5]. At the same time, there are chances that it may lose some amount of fine-grained temporal details because of performing the temporal averaging.

Advantages:

- Can encode overall shape and motion.
- Robust to noise.
- It can work as a reliable baseline descriptor.

3.2.2. Frame Difference Gait Image (FDGI)

FDGI captures the variance between consecutive frames in the silhouette sequence and emphasizes the temporal motion info, especially the changes in the limb positions over time. FDGI thus helps to capture the fine-grained details in gait cycle progression. The usage of FDGI features complements GEI by recovering the lost temporal dynamics, providing the model with motion-based cues that are vital for distinguishing individuals with similar static silhouettes. FDGI can thus clearly capture periodic movements and dominant gait rhythms. It improves temporal consistency and resilience to silhouette distortion [7,8].

Benefits:

- Effective capturing of periodic motion.
- Mostly unaffected by irregularities in silhouette sequences.
- Balances GEI by adding frequency-based temporal features.

3.2.3. Histogram of Oriented Gradients (HOG)

HOG is an interesting local spatial feature extractor that can capture edge orientation and gradient structures in gait silhouettes. It catches body contours and local variations in shape [9].

Benefits:

- Captures local texture and posture details.
- Complements GEI and FDGI by adding localized features.
- Strength against noise, illumination changes, and background clutter.
- Preservation of local appearance and shape, contributing to the texture details of the gait image.
- High discriminative power due to orientation and spatial distribution of gradients.

3.3. Novelty of Spatio-Temporal Feature Integration

The novelty of this proposed methodology is in the fusion of spatial (GEI, HOG) and frequency-based temporal (FDGI) features into a unified framework. While each of these features has individually been explored in past research, their joint integration in a temporally aware deep learning model represents a novel contribution.

The use of the CNN makes the model learn frame-wise spatial representations, at the same time preserving the sequential nature of the input. Followed by this, a multi-head attention mechanism is employed to capture long-range temporal dependencies across one complete gait cycle. This two-stage process makes sure that both local spatial and global temporal information are encoded in synergy, thereby improving recognition performance under cross-view, occlusion, and appearance-variant conditions.

Key novelty elements:

- Simultaneous exploitation of three distinct feature types.
- End-to-end trainable model with spatial and temporal modules.
- Novel fusion of complementary descriptors and deep temporal learning.

3.4. Model Architecture

1. Preprocessing: Silhouettes are extracted from video frames and are used to calculate GEI, FDGI, and HOG representations for each subject.
2. Time-Distributed CNN Layer: Each frame's feature map (GEI + FDGI + HOG stack) is passed through a CNN module which can apply shared weights to extract stable spatial features.
3. Temporal Attention Module: Output features from all time steps are fed into a multi-head self-attention mechanism that can capture sequence-level patterns.
4. Classification Layer: The aggregated feature vector is passed through a fully connected layer with softmax activation for classification.

3.5. Methodology

This section provides details of our end-to-end approach for a novel gait-based person identification system, which focuses on an automated preprocessing pipeline. Comprehensive feature extraction mechanisms, and a novel CNN–attention hybrid model that provides high accuracy and robustness, are employed in this model.

3.5.1. Automated Preprocessing Pipeline

In order to make sure that the quality and uniformity of input frames for gait recognition are kept high, we propose an automated preprocessing pipeline consisting of the following steps:

Step 1: Frame Extraction—Extract individual frames from gait video sequences using OpenCV library. Each frame is labeled with a unique person ID obtained from the video file name. Steps for the frame extraction are given in Algorithm 1.

Step 2: Normalization—Each frame is resized to 128 × 128 pixels. The frames are then converted to grayscale. The silhouette extraction procedure is applied on the grayscale images using background subtraction technique to highlight only the walking human.

Step 3: Organization—Frames are grouped based on extracted person identifiers to allow batch-wise training per subject.

After preprocessing:

- For each filename in output_dir:
    - Extract person_id using string parsing or regular expressions.
    - Group all frames belonging to the same person_id into separate folders or datasets.

3.5.2. Feature Extraction and Benefits

1. Gait Energy Image (GEI): GEI is widely used by people in research as a feature for gait recognition. GEI achieved this acceptance because of its ability to represent gait as a silhouette, which abstracts the person's walking pattern over a complete gait cycle [2,6]. It is calculated as the temporal averaging of silhouettes obtained from multiple frames in sequence. This results in a single binary image that encapsulates the motion over time.

$$\mathrm{GEI} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{S_t} \text{ for } \mathrm{t} = 1, 2, 3 \ldots \mathrm{T} \tag{1}$$

where

- S_t represents the silhouette at frame t.
- T is the total number of frames in the gait cycle.

The advantage of using GEI is that it can capture the overall motion pattern, supporting it to distinguish gait signatures, and is computationally efficient for real-time applications.

2. Frame Difference Gait Image (FDGI): FDGI is an advanced feature that gives importance to the temporal components of gait by converting temporal gait data into the frequency domain using the Fourier Transform. This highlights periodicities and rhythmic features of the gait motion which may not be visible in spatial domain representations such as GEI [7,8].

$$\text{The Fourier Transform } F(\omega) \text{of the gait signat } S(t) \text{ is given by } F(\omega) = \int_{-\infty}^{\infty} S(t) e(-j\omega t) \quad (2)$$

The Fourier coefficients, which are then used to represent gait. This allows us to focus on periodic gait patterns which are important for classification tasks.
FDGI improves the model's efficiency to recognize subtle temporal variations in gait, thereby improving robustness to variations like speed and stride length.

3. Histogram of Oriented Gradients (HOG): HOG is used to capture the distribution of gradient orientations in localized regions of an image [8]. This method effectively encodes the shape and movement characteristics of the human subject in terms of local edge and texture information. The HOG feature extraction uses Equation (3):
The FDGI feature is constructed by extracting significant temporal components from

$$HOG(x, y) = \sum_{i=1}^{n} Block_i(\sum_{p \in Block\ i} |\nabla I(p)|) \quad (3)$$

where $\Delta I(_p)$ is the gradient at pixel p;
Block refers to the set of pixels in the ith block;
N is the number of blocks in the image.
HOG is especially useful for gait recognition because it captures local details of body motion, such as joint movements, which is very important for distinguishing individual persons.

3.5.3. Spatio-Temporal Integration of Features

The proposed approach's novelty lies in the integration of both spatial and temporal features. Gait data is inherently temporal. Therefore, capturing the relationship between continuous frames is important for properly identifying gait patterns. So, by preparing the combination of GEI, FDGI, and HOG, we propose a method to jointly model the spatial and temporal dependencies of gait, which can make gait recognition more robust and more accurate.

We combine the GEI, FDGI, and HOG features along the spatial axis, resulting in a feature tensor that includes both local and global gait motion information. The combined feature C for each sequence of frames is represented using Equation (4):

$$C = [GE1_t, FDG1_t, HOG_tJ] \quad \text{for } t = 1\ 2,\ldots, T \quad (4)$$

where:

- GEI_t, $FDGI_t$, and HOG, are the respective features at time t.
- C is the concatenated feature tensor with shape (T, H, W, 3), where H and W are the height andwidth of the images, respectively.

This feature tensor contains both the temporal and spatial information at each time step, thus offering a comprehensive representation of a complete gait sequence.

Algorithm 1: Frame Extraction and Preprocessing

```
Input:
  video_path—Path to the video file
  output_dir—Directory where the output frames will be stored
  person_id—Identifier string for the subject in the video
Output:
  Grayscale, resized, and stored gait frames in the specified output directory
Begin
    Initialize cap ← VideoCapture(video_path)
    Set count ← 0
    Create directory output_dir if it does not already exist
    While cap is open do
        Read next frame: ret, frame ← cap.read()
        If ret = False then
            Break
        End If
        Convert frame to grayscale: gray ← cvtColor(frame, COLOR_BGR2GRAY)
        Resize gray to 128 × 88 pixels: gray_resized ← resize(gray, (128, 88))
        Extract silhouette using background subtraction: silhouette ←
        extract_silhouette(gray_resized)
        filename ← person_id + "_frame" + format(count, '04d') + ".png"
        Save silhouette to output_dir using filename
        Increment count ← count + 1
    End While
    Release video capture: cap.release()
End

Subroutine: extract_silhouette(frame)
This subroutine applies background subtraction to highlight the walking individual.
Begin
    Initialize background_subtractor ← cv2.createBackgroundSubtractorMOG2()
    silhouette ← background_subtractor.apply(frame)
    Return silhouette
End
```

3.5.4. Spatio-Temporal Integration of Features

The proposed approach's novelty lies in the integration of both spatial and temporal features. Gait data is inherently temporal. Therefore, capturing the relationship between continuous frames is important for properly identifying gait patterns. So, by preparing the combination of GEI, FDGI, and HOG, we propose a method to jointly model the spatial and temporal dependencies of gait, which can make gait recognition more robust and more accurate.

We combine the GEI, FDGI, and HOG features along the spatial axis, resulting in a feature tensor that includes both local and global gait motion information. The combined feature C for each sequence of frames is represented using Equation (4):

$$C = [GE1_t, FDG1_t, HOG_tJ] \quad \text{for } t = 1\ 2,\ldots, T \tag{4}$$

where:

- GEI_t, $FDGI_t$, and HOG, are the respective features at time t.
- C is the concatenated feature tensor with shape (T, H, W, 3), where H and W are the height andwidth of the images, respectively.

This feature tensor contains both the temporal and spatial information at each time step, thus offering a comprehensive representation of a complete gait sequence.

3.5.5. Modeling with CNN and Attention Mechanism

In order to use the temporal correlations, we use a Convolutional Neural Network (CNN) to extract spatial features from each of the frame data. This process is followed by an attention mechanism to model the inter-frame relationships. The CNN processes each frame independently through several convolutional layers:

$$F_{CNN} = \text{Conv2D}(F_{input}), F_{input} = [C_t] \quad (5)$$

where F_{Input} is the input feature tensor for each frame. This operation extracts spatial features from each frame independently, and the temporal dependencies across the frames are modeled using a multi-head attention layer (Transformer).

The output Y of the Transformer layer is given by the following:

$$Y = \text{MultiHeadAttention}(Q,K,V) \quad (6)$$

where Q, K, and V are the queries, keys, and values, respectively, derived from the CNN features.

Model Architecture and Training:

The combined CNN and Transformer model is trained using the softmax cross-entropy loss, where the output y_{pred} is the predicted class label and y_{true} is the true class label. The loss function is defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{n} y_{true,i} \log\left(y_{pred,i}\right). \quad (7)$$

y_{true} and y_{pred} are the true and predicted class labels, respectively.

We also use early stopping as a callback to prevent overfitting during training, with validation accuracy monitored to restore the best model weights.

3.5.6. Expected Contributions

The proposed approach seeks to address the limitations of traditional gait recognition systems by leveraging spatio-temporal feature fusion and deep learning architectures, offering the following:

- Enhanced robustness to variations in gait patterns.
- Improved generalization across different datasets.
- An effective framework for real-time gait recognition applications.

3.5.7. The Algorithm: Algorithm 2 Gives the Proposed Gait Recognition Steps

Algorithm 2: Gait Recognition Using Combined GEI, FDGI, and HOG Features with Transformer-Based Model

Input:

- *GEI, FDGI, and HOG image datasets for each person*
- *Sequence length SEQ_LENGTH*
- *Image size IMG_SIZE*

Output:

- *Trained model*
- *Accuracy metrics and confusion matrix*

Step 1: *Set Up Paths and Parameters*

1. *Define paths to the GEI, FDGI, and HOG image folders*
2. *Set IMG_SIZE = (64, 64) and SEQ_LENGTH = 5*

Step 2: *Load and Preprocess Data*

1. *For each person in the dataset:*
    - *Check if corresponding folders for GEI, FDGI, and HOG exist*
    - *Load, resize, normalize, and convert each image to grayscale*
    - *Concatenate GEI, FDGI, and HOG images horizontally*
    - *Form sequences of SEQ_LENGTH images*
2. *Append each sequence and corresponding label to sequences and labels*
3. *Convert sequences to numpy array and expand dimensions to match CNN input*
4. *Encode labels using LabelEncoder*

Step 3: *Split Dataset*

- *Use train_test_split() with stratification for creating X_train, X_val, y_train, y_val*

Step 4: *Define CNN + Transformer Model*

1. *Input shape: (SEQ_LENGTH, IMG_SIZE[0], IMG_SIZE[1]*3, 1)*
2. *CNN block (shared for all time steps):*
    - *Conv2D -> MaxPooling2D -> Conv2D -> MaxPooling2D -> Flatten -> Dense*
3. *Add MultiHeadAttention layer for self-attention across sequence*
4. *Add residual connection, normalization, global average pooling*
5. *Fully connected layers -> Output softmax layer with num_classes*

Step 5: *Compile Model*

- *Optimizer: Adam*
- *Loss: Sparse Categorical Crossentropy*
- *Metrics: Accuracy*

Step 6: *Train Model*

- *Use model.fit() with:*
    - *Batch size = 16*
    - *Epochs = 30*
    - *Validation data*
    - *EarlyStopping callback (patience = 5, restore best weights)*

Step 7: *Evaluate Model*

- *Track and display total training time*
- *Plot accuracy over epochs (train vs. val)*
- *Print final validation accuracy and loss*

Step 8: *Predict and Analyze Results*

- *Predict on validation set*
- *Calculate and print classification report*
- *Generate and display confusion matrix using heatmap*

End of Algorithm

3.6. Identified Gaps Leading to This Research

Although individual methods for gait recognition such as Gait Energy Images (GEI), Histogram of Gradients, etc., have been well researched, they exhibit limitations when

applied independently. Table 1 lists the advantages and disadvantages of various methods used.

Table 1. Advantages and limitations of methods from the literature.

Method	Advantages	Limitations
GEI	Captures global appearance and shape of walking silhouette	Loses temporal dynamics; sensitive to occlusion and clothing variations
CNN on GEI	Exploits spatial features effectively	Lacks temporal modeling; poor generalization across gait cycles
LSTM for sequences	Captures long-term dependencies	Suffers from vanishing gradients for long sequences
Attention-based Gait	Improves robustness by focusing on key features	Not optimized for multi-modal input (e.g., combining pose and motion info)

Gap Identified: There is a lack of an integrated framework that simultaneously captures appearance (GEI), structure (pose), and motion (trajectory) using attention-based spatio-temporal modeling.

3.7. Sequential Processing for Context Learning

The use of a CNN with a Transformer block allows the model to achieve the following:

- Apply self-attention to capture inter-frame dependencies, learning how motion evolves over time.
- Aggregate information across the sequence, enabling the model to utilize both short-term and long-term gait dynamics.
- The proposed architecture for gait-based person identification employs a hybrid deep learning model that integrates Convolutional Neural Networks (CNNs) with a Transformer encoder to effectively capture both spatial and temporal features inherent in gait sequences. The input to the model is a sequence of grayscale silhouette images with a fixed temporal window of five frames, each of size 64×192 pixels. A Time-Distributed CNN is applied to each frame independently, consisting of two convolutional layers with ReLU activations and max pooling for spatial down sampling, followed by a flattening layer and a dense layer that generates a 128-dimension feature embedding for each frame. This results in a sequence of five such embeddings that preserve the spatial characteristics of individual gait frames. These embeddings are then fed into a Transformer encoder block to model temporal dependencies and contextual relationships across the frame sequence. The Transformer employs a multi-head self-attention mechanism with four heads and a key dimension of 64, followed by residual connections and layer normalization to maintain stability and promote gradient flow. A position-wise feedforward network with ReLU activation and dropout further refines the temporal representations, again followed by residual connections and normalization to form the final output of the Transformer. To aggregate temporal information into a fixed-size representation, a global average pooling layer is applied, collapsing the sequence dimension and producing a unified 128-dimension vector. This vector is passed through a fully connected dense layer with ReLU activation and dropout for regularization and, finally, through a softmax output layer that maps the representation to the target classes. The architecture, by combining the spatial modeling power of CNNs with the sequence learning capabilities of Transformers, is well-suited for the complex task of gait recognition, capturing subtle variations in walking patterns across frames, and enabling robust person identification even under varying viewing conditions or walking styles.

This hybrid architecture enables the system to extract spatially rich representations while attending to the most informative temporal segments. Figure 1 shows a simple block diagram of the proposed methodology. Our novelty lies in the combination of motion trajectories, pose dynamics, and GEI for multichannel input and the integration of attention with temporal encoding to boost discriminative performance.

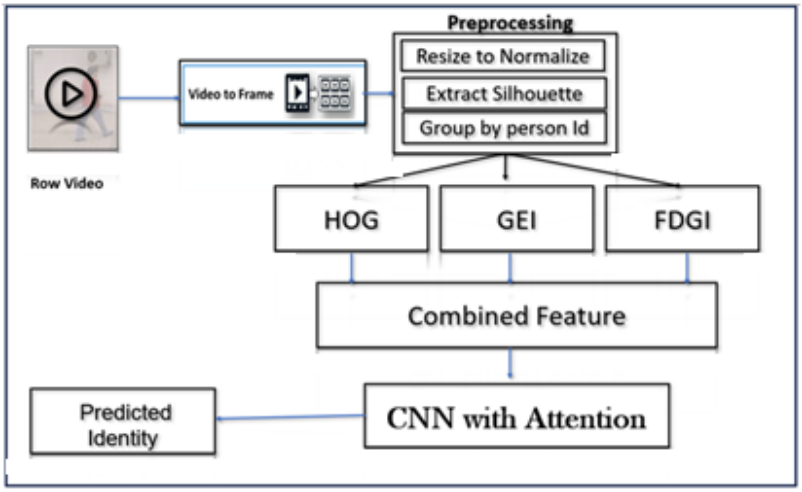

Figure 1. Block diagram for proposed methodology.

4. Results and Explanations

The proposed system was experimented to verify the performance while using various algorithms. In the system, a Convolutional Neural Network (CNN) integrated with a spatial attention mechanism to enhance the discriminative capability of combined HOG–FDGI–GEI features for the gait recognition was implemented.

The CNN model has consecutive convolutional and max pooling layers that extract hierarchical spatial features from grayscale input images of size 128 × 128. To improve the model's attention to salient areas of the input, a spatial attention module has been added after the second pooling layer. This module calculates the attention maps using average and max pooling over channels and then a convolutional layer to produce an attention weight map, which is multiplied with the input feature maps to highlight informative areas. The network is completed by fully connected layers and dropout regularization and outputs class predictions through a softmax layer. The model was trained using categorical cross-entropy loss and the Adam optimizer, with early stopping based on validation performance. This architecture demonstrated improved classification accuracy, as evidenced by validation metrics and confusion matrix analysis, indicating its effectiveness in capturing key gait patterns from the fused feature set.

For better attention of the model towards salient areas of the input, a spatial attention module has been added after the second pooling layer. This module calculates attention maps by using average and max pooling along channels, followed by a convolutional layer to produce an attention weight map, which is then element-wise multiplied with the initial feature maps to highlight informative areas. The network is completed with fully connected layers and dropout for regularization, and class predictions are output through a softmax layer.

The model was trained with categorical cross-entropy loss and the Adam optimizer and with early stopping on validation performance. This model showed better classification accuracy based on validation metrics as well as confusion matrix analysis, reflecting its ability to identify the most important gait patterns from the combined feature set. The default epochs considered were 30. The results are given in Table 2. Figure 2 shows the

confusion matrix for: FDGI and HOG, Figure 3 shows confusion matrix for GEI and FDGI and HOG. Figure 4 gives the comparison of accuracies and execution times for different feature combinatins.

Table 2. Accuracy received by various feature combinations tried.

Features Used	Algorithm	Accuracy	Execution Time (in s)
HOG	CNN + Attention	82.5	428.6
	MobileNet + Attention	80	529.6
FDGI	CNN + Attention	82.5	234.68
	MobileNet + Attention	77	326.3
GEI	CNN + Attention	80	608.8
	MobileNet + Attention	80	546.8
GEI + FDGI	CNN + Attention	92	406.27
	MobileNet + Attention	75	386.52
HOG + FDGI	CNN + Attention	95	227.15
	MobileNet + Attention	62	402.41
HOG + FDGI + GEI	CNN + Attention	97.5	202.41
	MobileNet + Attention	60	512.41

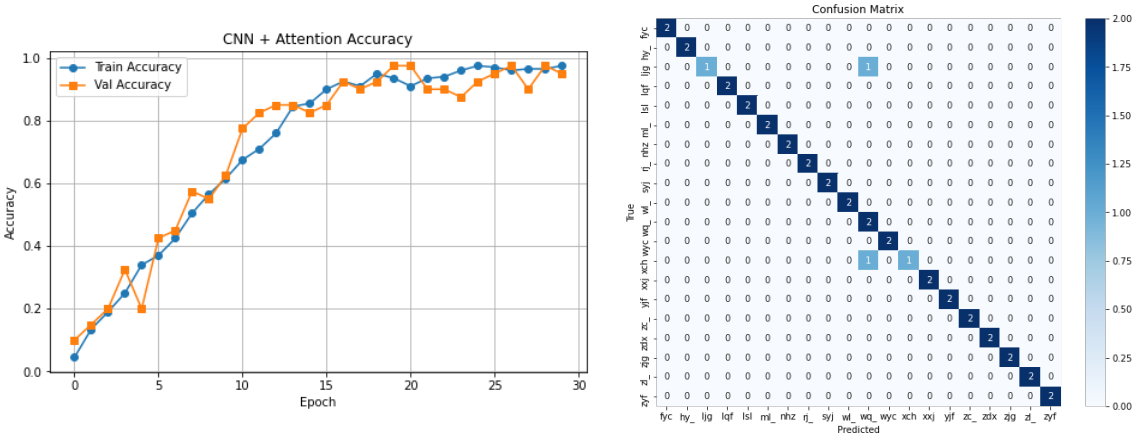

Figure 2. Accuracy (**left**) and confusion matrix (**right**). Features Used: FDGI and HOG.

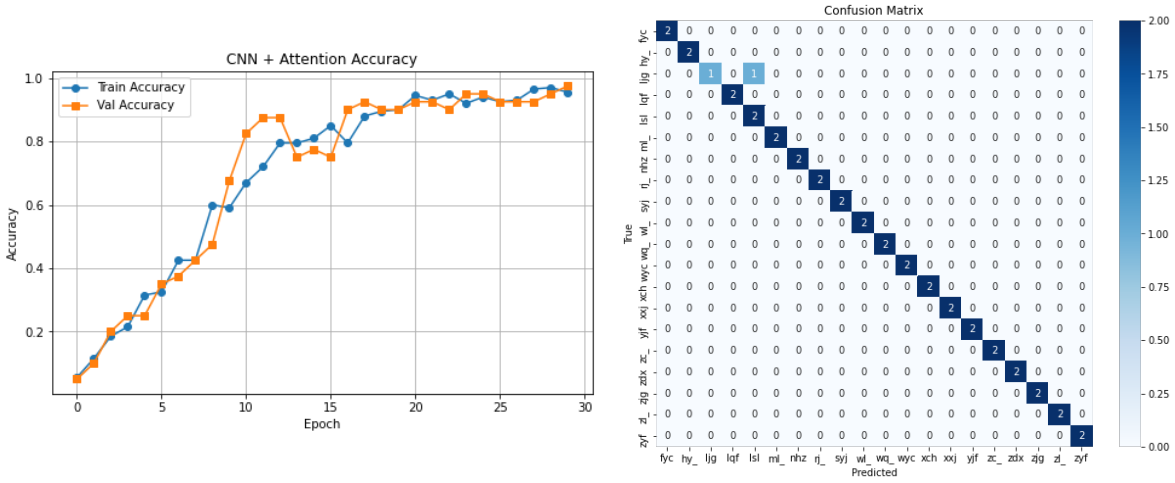

Figure 3. Accuracy (**left**) and confusion matrix (**right**). Features Used: GEI and FDGI and HOG.

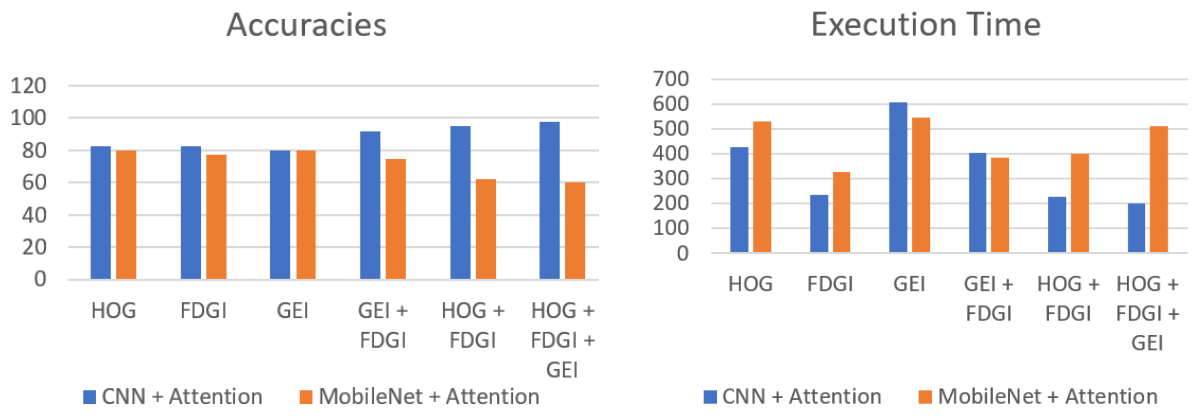

Figure 4. Comparison of accuracies (**left**) and execution time (**right**) when different feature combinations were used in CNN + Attention and MobileNet + Attention.

5. Conclusions

The work presented in this paper is an effective gait recognition framework for human identification that brings together spatial and temporal features and uses an attention mechanism. Experiments were performed with GEI, HOG, and FDGI. The blend of these features greatly enhances the model's skill to capture discriminative gait patterns on various viewpoints. Experimental evaluations show the robustness and improvement accuracy of the proposed approach. Future work can be performed on extending the model to handle real-time gait recognition, cross-dataset generalization, and performance in unconstrained environments such as occlusions or varying walking speeds.

Author Contributions: Conceptualization, K.T.T. and K.P.P.; methodology, K.T.T.; software, K.T.T.; validation, K.T.T. and K.P.P.; writing—original draft preparation, K.T.T.; writing—review and editing, K.P.P.; visualization, K.T.T.; supervision, K.P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable. This study does not require ethical clearance.

Informed Consent Statement: Not applicable as data was not taken from human subjects.

Data Availability Statement: No data is generated while doing the work specified in the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Han, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 316–322. [CrossRef] [PubMed]
2. Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR), Hong Kong, China, 20–24 August 2006; pp. 441–444. [CrossRef]
3. Boulgouris, N.V.; Hatzinakos, D.; Plataniotis, K.N. Gait recognition: A challenging signal processing technology for biometric identification. *IEEE Signal Process. Mag.* **2005**, *22*, 78–90. [CrossRef]
4. Muramatsu, H.; Makihara, Y.; Yagi, Y. View transformation model incorporating quality measures for cross-view gait recognition. *IEEE Trans. Cybern.* **2016**, *46*, 1602–1615. [CrossRef] [PubMed]
5. Kale, A.; Roy-Chowdhury, A.K.; Chellappa, R. Fusion of gait and face for human identification. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, QC, Canada, 17–21 May 2004; Volume 5, pp. V-901–V-904.
6. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.

7. Chao, Y.; Xu, D.; Xu, C. GaitSet: Regarding gait as a set for cross-view gait recognition. *AAAI Conf. Artif. Intell.* **2019**, *33*, 8126–8133. [CrossRef]
8. Thomas, K.T.; Pushpalatha, K.P. Deep Learning-based Gender Recognition Using Fusion of Texture Features from Gait Silhouettes. In *Data Science and Security: Proceedings of IDSCS 2022*; Lecture Notes in Networks and Systems; Springer: Singapore, 2022; Volume 462. [CrossRef]
9. Lam, T.H.; Lee, K.H.; Siu, W.C. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognit.* **2011**, *44*, 973–987. [CrossRef]
10. Wang, L.; Tan, T.; Ning, H.; Hu, W. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1505–1518. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.