electronics

MDPI

Article

Multimodal Low Resolution Face and Frontal Gait Recognition from Surveillance Video

Sayan Maity *, Mohamed Abdel-Mottaleb and Shihab S. Asfour

Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146-0620, USA; mottaleb@miami.edu (M.A.-M.); sasfour@miami.edu (S.S.A.)

* Correspondence: s.maity1@umail.miami.edu

Abstract: Biometric identification using surveillance video has attracted the attention of many researchers as it can be applicable not only for robust identification but also personalized activity monitoring. In this paper, we present a novel multimodal recognition system that extracts frontal gait and low-resolution face images from frontal walking surveillance video clips to perform efficient biometric recognition. The proposed study addresses two important issues in surveillance video that did not receive appropriate attention in the past. First, it consolidates the model-free and model-based gait feature extraction approaches to perform robust gait recognition only using the frontal view. Second, it uses a low-resolution face recognition approach which can be trained and tested using low-resolution face information. This eliminates the need for obtaining high-resolution face images to create the gallery, which is required in the majority of low-resolution face recognition techniques. Moreover, the classification accuracy on high-resolution face images is considerably higher. Previous studies on frontal gait recognition incorporate assumptions to approximate the average gait cycle. However, we quantify the gait cycle precisely for each subject using only the frontal gait information. The approaches available in the literature use the high resolution images obtained in a controlled environment to train the recognition system. However, in our proposed system we train the recognition algorithm using the low-resolution face images captured in the unconstrained environment. The proposed system has two components, one is responsible for performing frontal gait recognition and one is responsible for low-resolution face recognition. Later, score level fusion is performed to fuse the results of the frontal gait recognition and the low-resolution face recognition. Experiments conducted on the Face and Ocular Challenge Series (FOCS) dataset resulted in a 93.5% Rank-1 for frontal gait recognition and 82.92% Rank-1 for low-resolution face recognition, respectively. The score level multimodal fusion resulted in 95.9% Rank-1 recognition, which demonstrates the superiority and robustness of the proposed approach.

Keywords: multimodal biometrics; low resolution face; super resolution; frontal gait

Citation: Maity, S.; Abdel-Mottaleb, M.; Asfour, S.S. Multimodal Low Resolution Face and Frontal Gait Recognition from Surveillance Video. *Electronics* **2021**, *10*, 1013. https://doi.org/10.3390/electronics10091013

Academic Editor: Athanasios Voulodimos

Received: 8 March 2021
Accepted: 20 April 2021
Published: 24 April 2021

1. Introduction

The importance of identifying and monitoring the activity of registered offenders using video surveillance footage has been proven effective on several occasions, e.g., identifying the Boston bombing suspects, to lead the detectives in the right direction. However, the quality of the video data acquired by the surveillance system poses challenges. The primary causes of poor image quality recorded in most digital video surveillance systems are low resolution, excessive quantization, and low frame rate. Moreover, high-resolution video surveillance systems require excess storage space. These factors result in low-resolution biometric data, e.g., face images, obtained from the video surveillance clips collected using the existing video surveillance systems.

In this paper, we propose a solution for accurate human identification from low-resolution video surveillance footage by combining gait recognition and low-resolution (LR) face recognition. The proposed system, shown in Figure 1, is a fully automatic platform

which first extracts the frontal gait silhouettes and low-resolution face images from the frontal walking video surveillance clips. Then, it obtains the feature vectors from the pre-processed frontal gait silhouettes, and the low-resolution face images. Later the feature vectors are used to train two separate classifiers to perform the frontal gait recognition, and low-resolution face recognition. Finally, the individual recognition results are fused through score level fusion. Given a test surveillance video clip of a subject walking towards the camera, first the gait features and LR face image features are extracted, later nearest neighbor classifiers are used to separately obtain the Rank-1 frontal gait recognition and LR face recognition results. Finally, score level fusion is performed to fuse the individual recognition results.

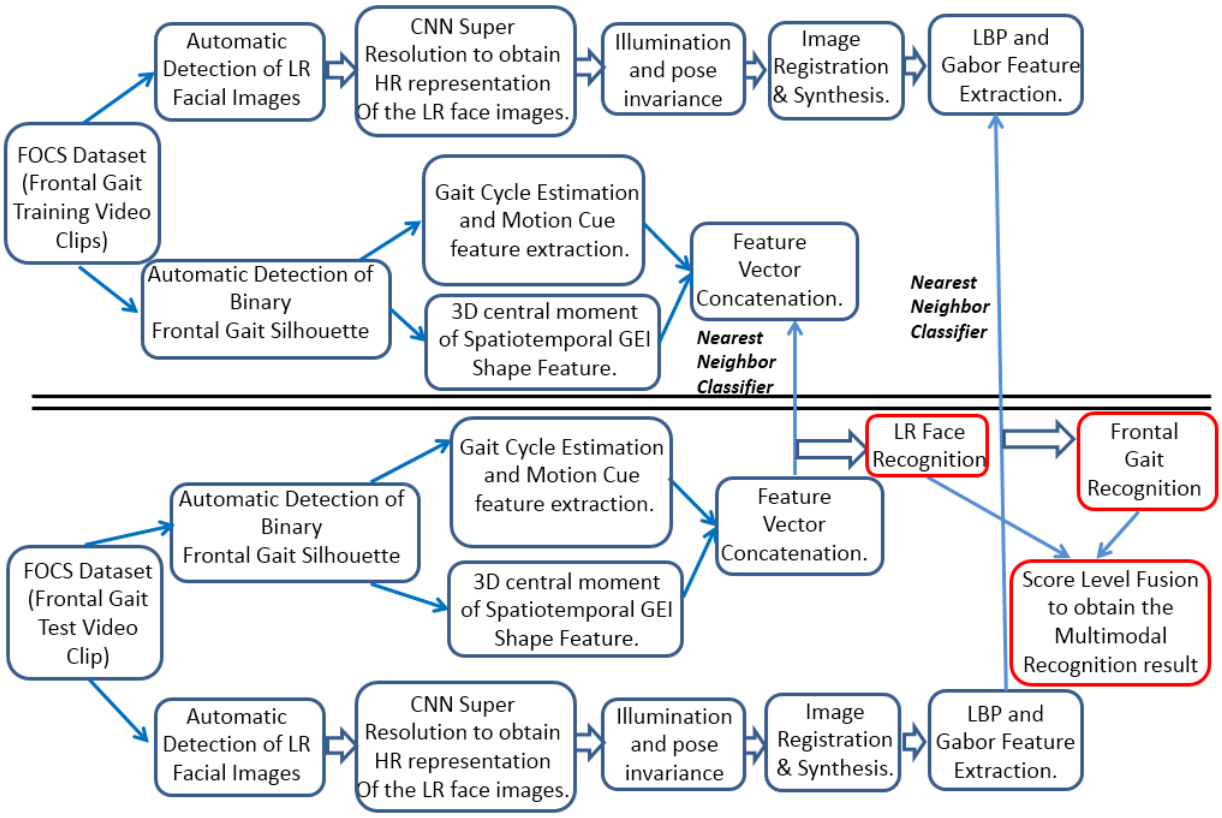

Figure 1. System block diagram: multimodal biometrics recognition from video survillience data.

Due to the unavailability of proper datasets for multimodal face and gait recognition, the proposed studies in the literature were evaluated only on databases with a small set of subjects [1,2]. Moreover, majority of the approaches in the literature use lateral gait view, or use camera calibration, or even require multiple cameras for capturing multiple gait views to perform gait recognition. Gait cycle detection is critical for gait feature extraction and can be efficiently detected from the lateral gait view. Majority of the studies on gait recognition [3,4] perform the gait cycle approximation using various heuristic approaches from the biomechanics literature, which incorporate significant estimation error when applied on subjects with a wide-range of walking speed in large databases. In a practical situation, a system which estimates the gait parameters from a single view, without depending on the subject's pose or on camera calibration, is more realistic. We propose an efficient gait recognition technique through a robust gait cycle detection using frontal gait video clips. Previous studies of gait recognition in the literature apply either model-free or model-based approaches for feature extraction. We incorporate both the model-free (i.e., average walking speed for a subject, which is determined using the average number of frames, detected from the video, for the gait cycles) and the model-based (scale

and translation invariant 3D moments for shape feature extraction) gait feature extraction approaches for robust identification.

A considerable amount of literature has been published on low-resolution face recognition. The majority of these studies use high resolution (HR) images/video to synthetically generate the corresponding low-resolution counterpart. Then a mapping function is obtained between the high- and low-resolution image pair. In this paper, we only use low-resolution face information obtained from the video surveillance data to train and later test the performance of the proposed low-resolution face recognition algorithm. It is evident from the experiment results that the proposed framework allows the system to learn the high-resolution mapping function from the low-resolution images, results in considerably higher classification accuracy by maximizing the signal-to-noise ratio. To the best of our knowledge, the proposed approach is the first fully automatic multimodal recognition framework using LR face images and frontal gait silhouettes from surveillance video clips. Compared to other studies, the performance is evaluated on a relatively large dataset.

2. Related Work

Even though research in face recognition has been active for the past few decades [5–9], the topic of low-resolution [10] face recognition has only recently received much attention, for long distance surveillance applications, to recognize faces from small size or poor quality images with varying pose, illumination, and expression. Although the state-of-the-art face recognition accuracy using data collected in constrained environments is satisfactory, the recognition performance in real world applications such as video surveillance is still an open research problem, primarily due to low-resolution (LR) images [11] and variations in pose, lighting conditions, and facial expressions.

Gait recognition [12,13] is a well proven biometric modality, which can be used to identify a person remotely through inspecting their walking patterns. However, gait recognition has some subtle shortcomings: it can be affected by the dressing attire, carrying large objects, etc. Moreover, the physical state, such as injuries, can also affect a person's walking pattern. Majority of the proposed gait recognition techniques [14,15] employ multi-view gait recognition to overcome the viewing angle transformation problem and to improve the recognition accuracy.

2.1. Low-Resolution Face Recognition

The literature on low-resolution face recognition can be categorized into three broad classes:

(1) **Mapping into unified feature space**: In this approach, the HR gallery images and LR probe images are projected into a common space [16]. However, it is not straight forward to find the optimal inter-resolution (IR) space. Computation of two bidirectional transformations from both HR and LR to a unified feature space usually incorporates noise.

(2) **Super-resolution**: Many researchers used up-scaling or interpolation techniques, such as cubic interpolation on the LR images. Conventional up-scaling techniques usually are not good for the images with relatively lower resolution. However, super-resolution [17,18] methods can be utilized to estimate HR versions of the LR ones to perform efficient matching.

(3) **Down-scaling**: Down-sampling techniques [11] can be applied on the HR images followed by comparison with the LR image. However, these techniques are poor in performance for solving LR problem, primarily because the downsampling reduces the high-frequency information which is crucial for recognition.

Due to the challenges and importance for real world applications, low-resolution (LR) face recognition has gradually become an active research area of Biometrics in recent years. Ren et al. [16] proposed a novel feature extraction method for face recognition from LR images, i.e., coupled kernel embedding, where a unified kernel matrix is constructed by concatenating two individual kernel matrices obtained, respectively, from HR and LR images. Sumit et al. [19], proposed an approach for building multiple dictionaries of different

resolutions and after identifying the resolution of the probe image a reconstruction error based classification is obtained. A very low resolution (VLR) face recognition technique is proposed in [11] with a resolution lower than 16×16 by modeling the relationship between HR and VLR images using a piecewise linear regression technique. A super-resolution-based LR face recognition technique is proposed by Yu et al. [20], where an LR face image is split into different regions based on facial features and the HR representation of each section is learned separately. Jia et al. [21] proposed a unified global and local tensor space representation, to obtain the mapping functions to acquire the HR information from the LR images to perform efficient LR face recognition.

2.2. Gait Recognition

The first step of gait recognition is background subtraction. The feature extraction techniques in the literature [22] can be categorized broadly in two classes:

(1) **Model-free approaches**: In the model-free gait representation [23], the features are composed of a static component, i.e., size and shape of a person, and a dynamic component, which portrays the actual movement. Examples of static features are height, stride length, and silhouette bounding box. Whereas dynamic features can include frequency domain parameters like frequency and phase of the movements.
(2) **Model-based approaches**: In the model-based gait representation approaches [13,24] we need to obtain a series of static or dynamic gait parameters via modeling or tracking the entire body or individual parts such as limbs, legs, and arms. Gait signatures formed using these model parameters are utilized to identify an individual.

Model-free approaches are usually insensitive to the segmentation quality and less computationally expensive compared with the model-based approaches. However, the model-based approaches are usually view-invariant and scale-independent compared with the model-free counterpart.

To obtain the gait signature utilizing a sequence of gait silhouettes, Davis et al. [25] proposed the motion-energy image (MEI) and motion-history image (MHI) which transform the temporal sequence of silhouettes to a 2D template for gait identification. Later, Han and Bhanu [13] adopted the idea of motion-energy image (MEI) and proposed the gait energy image (GEI) for individual recognition using gait images. Frequency analysis of spatio-temporal gait signals is used by researchers to model the periodical gait cycles. Lee et al. [23] proposed a model-free approach to first divide the gait silhouette into seven regions and align them with ellipses, later apply Fourier Transform on the fitted ellipses to extract the magnitude and phase components for classification. Goffredo et al. [26] proposed a k-nearest neighbor classifier (k-NN) for front-view gait recognition where the gait signature is composed of shape features extracted from sequential silhouettes.

2.3. Multimodal Face and Gait Recognition

The fusion of face and gait modalities have recently received significant attention [27], mainly motivated by their impact on security related applications. The fusion of the two modalities has been used in the literature to obtain more robust and accurate identification. The fusion can be performed at the feature/sensor level, the decision level, or the matching score level.

In [28], features from high-resolution profile face images and features from gait energy images are extracted separately and combined at the feature level, later the fused feature vector is normalized and used for multimodal recognition. The experimental results on a database of video sequences for 46 individuals demonstrate that the integrated face and gait features result in a better performance than the performance obtained from the individual modalities. Shakhnarovich et al. [1] proposed a view normalized multimodal face and gait recognition algorithm and evaluated it on a dataset of 26 subjects. First, the face and the gait features are extracted from multiple views and transformed to the canonical pose frontal face and the profile gait view, later the individual face and gait recognition results are combined at the score level. In [2], a score level fusion of face and

gait images from a single camera view is proposed and tested on an outdoor gait and face dataset of 30 subjects. The results of a view-invariant gait recognition algorithm, and a face recognition algorithm based on sequential importance sampling are fused in a hierarchical and holistic fashion. Geng et al. [29] proposed a context-aware multi-biometric fusion of gait and face which dynamically adapts the fusion rules to the real-time context and respond to the changes in the environment.

3. Materials and Methods

To perform multimodal biometric recognition, we need to detect the face and the gait silhouette from the surveillance video clips. The surveillance video clips are captured by a static video camera which records the frontal view of a walking person. The subjects start walking form a distance directly approaching the camera. We extract both the frontal gait silhouettes from the sequence of video frames and the low-resolution frontal face images, as explained below.

We adopted the fast object segmentation method proposed by Papazoglou et al. [30] for segmenting the foreground silhouette from the background. The fast object segmentation is fast, fully automatic, and has minimal assumptions about the motion of the foreground. Therefore, it performs efficiently in cases of unconstrained settings, presence of rapidly moving objects, arbitrary object motion and appearance change, and non-rigid deformations and articulations. The fast object segmentation technique first produces a rough estimate of the pixels that are inside the object, based on motion boundaries using optical flow obtained from pairs of subsequent frames [31,32]. In the second step, a spatio-temporal extension of GrabCut [33,34] technique is used to bootstrap an appearance model based on the initial foreground estimate, and refine it by integrating information over the entire video sequence. An example of segmented silhouettes from different frames, using fast object segmentation [30], is shown in Figure 2, which shows accurate segmentation by isolating the silhouette from its reflection on the shiny floor.

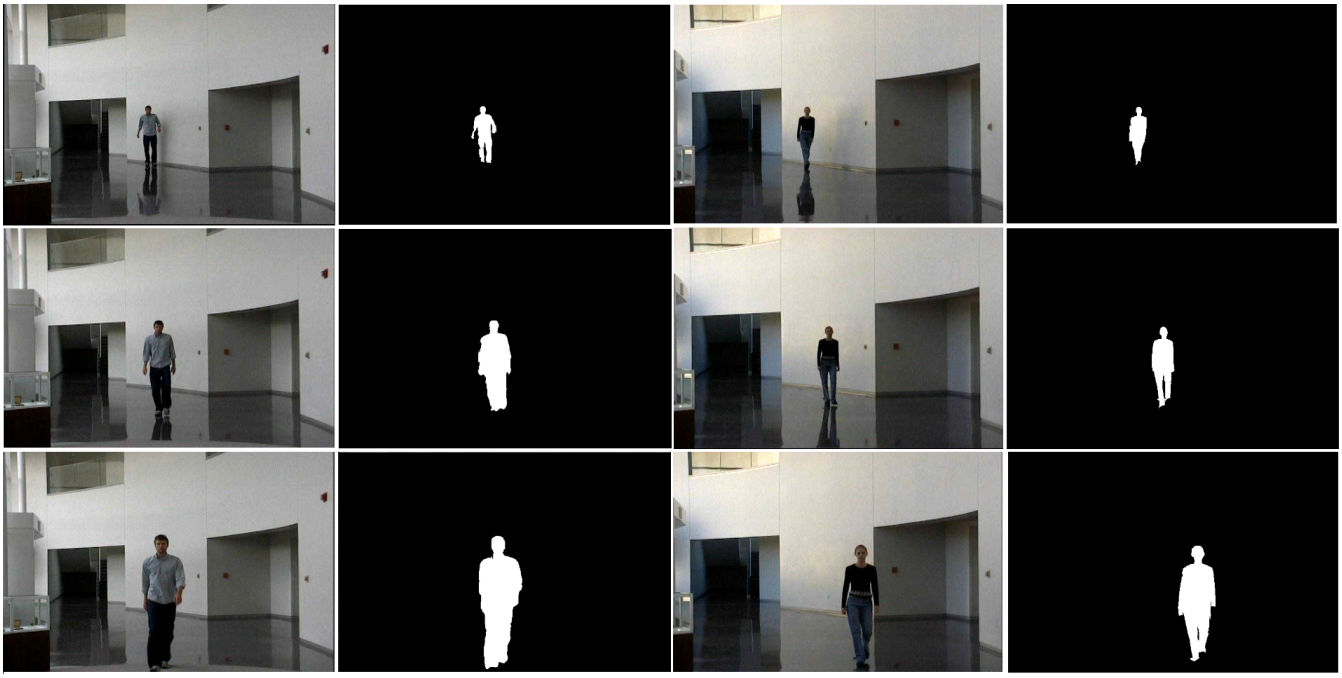

Figure 2. Frontal gait silhouette segmentation.

To automatically detect the low-resolution frontal face images in the surveillance video clips, we adopt the Adaboost object detection technique, proposed by Viola and Jones [35]. The algorithm is trained to detect low-resolution frontal faces using manually cropped frontal face images from the color FERET [36] database. By using the trained detector, we

can detect low-resolution faces in the video frames. The trained detector is applied to the entire video sequence to detect the LR frontal faces. An example of the detection results from a surveillance video clip is shown in Figure 3.

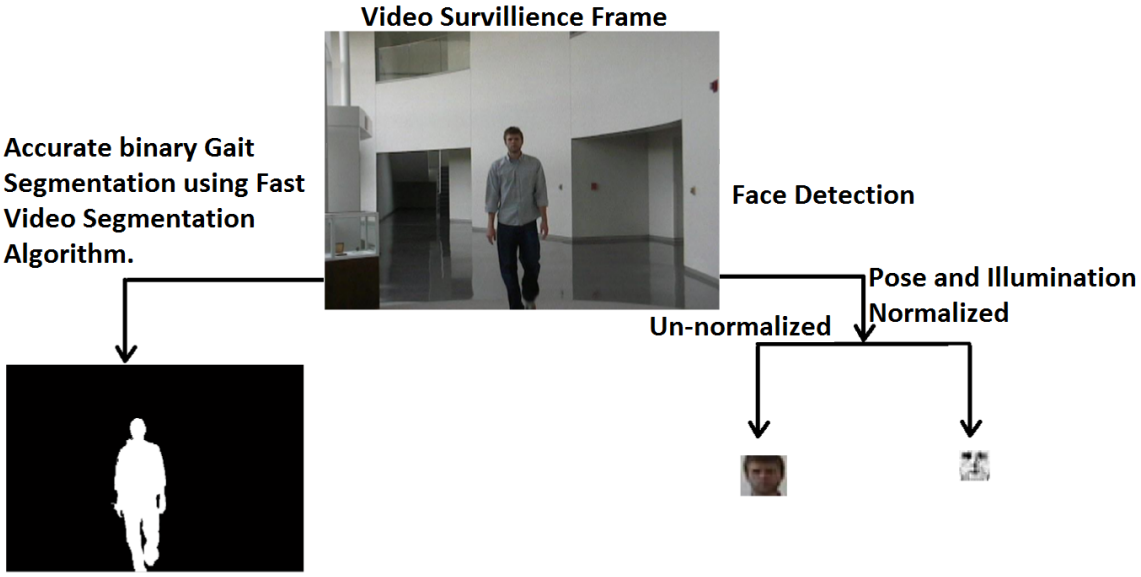

Figure 3. Gait and low-resolution face extraction.

Existing studies in the literature [37,38] suggests that human periodic movement speeds and patterns are similar in repeated trials of the same subject. We have incorporated both model-free and model-based feature representation of the segmented silhouettes to obtain accurate and efficient gait recognition. Identification of the gait cycle, using the frontal gait video, is proposed to compute average movement speed for efficient model-free gait recognition. Moreover, model-based gait energy image (GEI) [13] features are also extracted to perform view-invariant and scale-independent gait recognition.

In the following subsections, we described the proposed method of gait cycle identification to compute average movement speed and 3D moments from the spatio-temporal GEI shape feature, using the segmented silhouettes.

3.1. Gait Cycle Identification Using Frontal Gait

3.1.1. Gait Feature Representation

In this section, we first define the gait cycle and then describe the proposed approach to identify gait cycle using only frontal gait information.

The gait cycle [37] can be defined as the time interval between two successive occurrences of the repetitive phases while walking. The gait cycle involves two principal stages: the stance phase and the swing phase. The stance phase occupies 60% of the gait cycle, while the swing phase occupies only 40%, as explained in Figure 4. The stance phase consists of Initial Contact, Loading Response, Midstance, Terminal Stance, and Pre-swing. Whereas, the swing phase is composed of Initial Swing, Mid Swing, and Terminal Swing. Stance phase begins with the heel strike—this is the moment when the heel begins to touch the ground, but the toes do not yet touch. We can see from Figure 4, during the Stance phase, in the Midstance position, the difference between the lower points (or pixel locations) of the two limbs is maximized. Similarly, in the Midswing position of the swing phase, in-between the Initial Swing and the Heel Strike, the distance between the lower points of the two limbs is maximized. Whereas, during the Terminal-swing through Loading Response stages, the distance between the lowest white pixel of the two limbs is minimized.

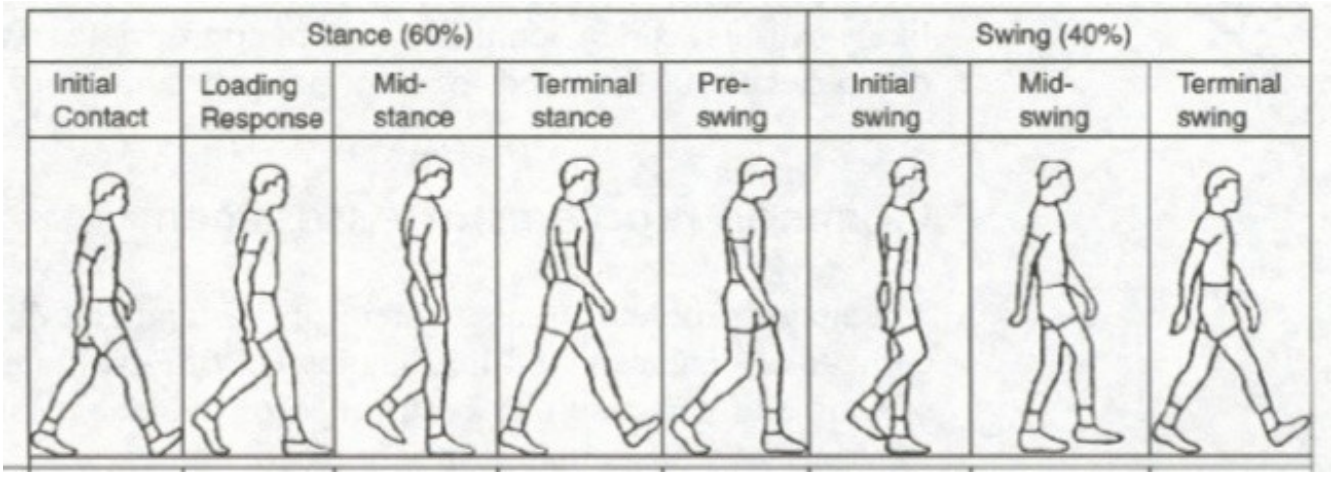

Figure 4. Gait cycle definition.

Following this specific attribute of the gait cycle, we can analyze gait cycle from the frontal silhouette. In Figure 5, we can see that in the silhouette bounding box of frames 138 and 152 the difference between the lowest white pixel of the two limbs is maximum which indicates the successive events of the Midstance through Midswing phase. Moreover, in the silhouettes of frame 144 and frame 158, the difference between the lowest white pixel of the two limbs is minimum, which signifies the successive events of Pre-swing through Terminal swing. Therefore, we can identify the entire gait cycle from the sequence of frontal gait silhouettes starting from the Initial Contact (frame 135) through Terminal swing (frame 158) in Figure 5.

Figure 5. Gait cycle estimation.

Identifying the gait cycles from the gait video is usually the initial step in gait analysis for separating the periodic occurrences of the walking sequence. Majority of the techniques [13,39] in the literature perform the detection of the gait cycle using profile gait view or multiple gait views due to the ease of discrimination of different gait phases as described earlier. As per biological studies [40,41] of the human gait cyclic phases during walking, the body pose changes periodically and the upper and lower limbs move symmetrically. Since the width and height of the bounding box of the binary silhouette directly depends on the limb's fluctuation, we represent the gait fluctuation as a periodic function which depends on the silhouette's width and height over time. In a frontal gait video, as the subject is moving towards the surveillance camera, the silhouette's height and width will be increasing in the later frames compared with the earlier ones. To compensate for these scale variations, we normalized [42] the width and height of the silhouette bounding box. Based on the theoretical premise of the gait cycle and the experimental observations using the frontal gait video clips, we propose a gait cycle identifier which represents the periodic motion and cyclic phases as follows:

$$\begin{aligned} GC_{identifier}(f_t) = 0.5 * [H_{norm}(f_t) * W_{norm}(f_t) + \\ \{||\text{lowest left limb pixel} - \text{lowest right limb pixel}||\}/H(f_t)], \end{aligned} \quad (1)$$

where $GC_{identifier}(f_t)$ is the variable that represents the gait cycle phase for the tth frame (f_t), $H_{norm}(f_t)$ and $W_{norm}(f_t)$ are the silhouette's bounding box height and width for the tth frame after normalization to compensate for the scale variations. The second term in Equation (1) is the normalized difference between the lowest white pixels of the two limbs, where $H(f_t)$ is the height of the t-th frame. The multiplier 0.5 is used to normalize the value of the gait cycle identifier variable. The plot of the $GC_{identifier}(t_f)$ against the sequence of frames is shown in Figure 6.

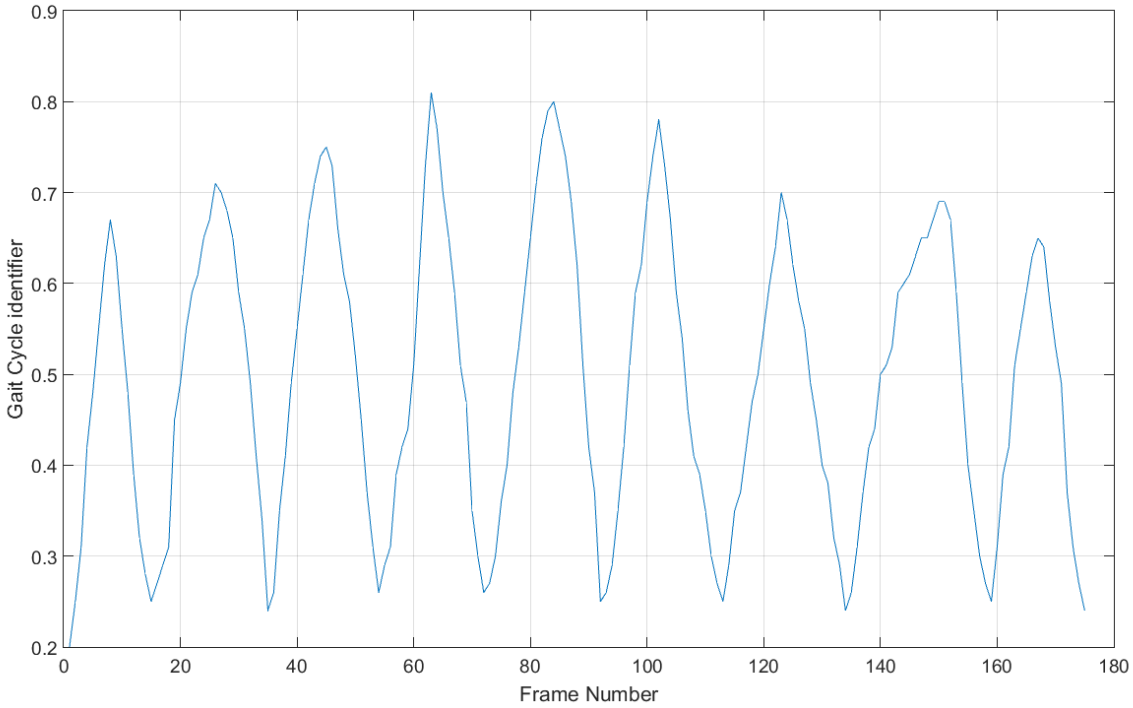

Figure 6. Gait cycle plot using GC identifier.

3.1.2. Three Dimensional Moments from the Spatio-Temporal Gait Energy Image

After the silhouettes are segmented from each of the video frames, their heights are first normalized with respect to the frame height. The average silhouette image or the gait energy image (GEI) [13] represents the principal shape of the human silhouette and its

change over a sequence of frames in a gait cycle. A pixel with higher intensity value in the GEI indicates that the human body was present more frequently at this specific position. Equation (2) is used for obtaining the pixel values of the GEI:

$$G(x, y) = \frac{1}{F} \sum_{t=1}^{F} B_t(x, y), \tag{2}$$

where t stands for the temporal frame number from which the silhouette is obtained, F is the total number of frames in a complete gait cycle, $B_t(x, y)$ stands for the binary silhouette. Spatio-temporal GEI or the periodic gait volume $V(x, y, n)$ is obtained from the GEIs computed using the gait cycles in a gait video clip, where n represents the gait cycle number.

Even though GEI suffers from some information loss of the details, it has numerous benefits compared with the representation of binary silhouettes as a temporal sequence. Since GEI is the average of a sequence of silhouettes, it is not very sensitive to errors in the silhouette segmentation in the individual frames. The robustness of the GEI is improved by discarding pixels with the energy values lower than a predefined threshold.

Shape analysis is a complex problem due to the presence of noise, and in certain cases, variations between shapes result in significant changes in the measured feature values. To recognize objects from their shape, features such as eccentricity, Moments, Euler number, compactness, and convexity are widely used in the literature [43]. Moments or central moments are used as quantitative measures for shape description [44]. Hu et al. [44] derived a set of moment invariants for various geometric shapes. Moments are widely used in various complex shape-based object recognition [45] due to the fact that they are invariant to orientation.

Three-dimensional raw moments for the s-temporal GEI or periodic gait volume for each gait cycle can be represented as:

$$3DMoment_{p_1p_2p_3} = \sum_{x \in \mathbf{x}} \sum_{y \in \mathbf{y}} \sum_{n \in \mathbf{n}} x^{p_1} \cdot y^{p_2} \cdot n^{p_3} \cdot V(x, y, n), \tag{3}$$

where order (O) of 3D moments can be represented as: $O = p_1 \cdot p_2 \cdot p_3$. For any translation, e.g., (a, b, c), of the 3D coordinates of the center of mass of the object, the change in the three dimensional moments $3DMoment_{p_1p_2p_3}$ can be represented as:

$$\overline{3DMoment}_{p_1p_2p_3} = \sum_{x \in \mathbf{x}} \sum_{y \in \mathbf{y}} \sum_{n \in \mathbf{n}} (x + a)^{p_1} \cdot (y + b)^{p_2} \cdot (n + c)^{p_3} \cdot V(x, y, n), \tag{4}$$

When the center of mass $(\bar{x}, \bar{y}, \bar{n})$ is at the origin, the raw moments and the central moments are the same. Thus, the central moment $\mu_{p_1p_2p_3}$ can be represented by replacing a, b, c with the mean value of x, y, n, respectively:

$$\mu_{p_1p_2p_3} = \sum_{x \in \mathbf{x}} \sum_{y \in \mathbf{y}} \sum_{n \in \mathbf{n}} (x - \bar{x})^{p_1} \cdot (y - \bar{y})^{p_2} \cdot (n - \bar{n})^{p_3} \cdot V(x, y, n). \tag{5}$$

Here,

$$\bar{x} = \frac{m_{100}}{m_{000}}; \quad \bar{y} = \frac{m_{010}}{m_{000}}; \quad \bar{n} = \frac{m_{001}}{m_{000}}, \tag{6}$$

where m_{000} is the zeroth spatial moment, and m_{100}, m_{010}, and m_{001} are the x, y, and n components of the first spatial moment, respectively. The pixel on the pereodic gait volume, e.g., $[x_j(n), y_j(n)]$, of $V(x, y, n)$ is the jth point that belongs to the n-th gait cycle. Hence, the 3D central moment of the spatio-temporal GEI or the periodic gait volume can be represented as:

$$\mu^{GEI_{vol}}_{p_1p_2p_3} = \sum_{n \in \mathbf{n}} \sum_{j=1}^{P(n)} (x_j(n) - \bar{x})^{p_1} \cdot (y_j(n) - \bar{y})^{p_2} \cdot (n - \bar{n})^{p_3}, \tag{7}$$

where P(n) is the total number of pixels on the periodic gait volume for gait cycle n.

Following the method mentioned in the previous two sections, we obtained the scale and translation invariant three-dimensional moments of the periodic gait volume ($\mu_{P_1P_2P_3}^{GEI_{vol}}$). Additionally, the average number of frames in the gait cycles identified using the frontal walking video clip is used as the average movement speed of the subject. We used these two components together to obtain the gait signature used for classifying the subjects through gait recognition.

3.2. Low-Resolution Face Feature Representation

In this section, we describe the proposed algorithm for low-resolution face recognition from surveillance video clips. The description of the components used in the algorithm are detailed in the subsequent sections. Here the algorithm refer to the proposed technique in the manuscript.

3.2.1. Super-Resolution

Super-resolution (SR) [17,18] is a class of image processing algorithms, used to enhance the resolution of low-resolution images. SR algorithms can be used to enhance the resolution of an image from single or multiple low-resolution images. Interpolation techniques such as nearest neighbor, bilinear and cubic convolution are widely used for SR processing of the LR images in the literature.

The two key components of a digital imaging system are the sensor and the lens, those introduce two types of image degradation, specifically optical blur and limitation on the highest spatial frequency that can be recorded. The sensor is constructed from a finite number of discrete pixels which results in the presence of so-called aliased components in the sensor output. These correspond to high spatial-frequency components in the scene that are higher than frequencies that the sensor can handle and should not normally be present in the output. These are the key components used by the SR algorithms to obtain the HR representation. The available SR algorithms can be categorized broadly into two major classes: reconstruction-based SR and recognition-based SR. The reconstruction-based methods are suitable for synthesizing local texture resulting in better visualization and do not incorporate any specific prior information. However, recognition-based SR [17,18] algorithms try to detect or identify certain pre-configured patterns in the low resolution data.

The recognition-based SR algorithms [17] learn a mapping correspondence between low and high resolution image patches from the training LR and HR images, which can be directly applied to a test LR image to construct the HR counterpart. In the training phase densely overlapping patches are cropped from the low-resolution and high-resolution image pair. Followed by jointly training two dictionaries for the low- and high-resolution image patches by enforcing the similarity of sparse representation for each image pair. Given the trained LR and HR dictionaries and a test LR image, the algorithm obtains its HR representation in three steps. First, densely overlapping patches are cropped from the LR input image and pre-processed (i.e., normalization). Second, the sparse coefficients obtained from the LR dictionary for the LR test image patches are passed into the high-resolution dictionary for reconstructing the high-resolution patches. Finally, the overlapped HR reconstructed patches are aggregated (i.e., weighted averaging) to produce the final output.

Convolutional neural network (CNN) [46] was developed several decades ago and deep Conv Nets [47] have recently been popular among researchers primarily due to its success in image classification. CNN is a specific artificial neural network topology, that is inspired by biological visual cortex, formed by stacking multiple stages of feature extractors. CNN have also been used successfully for other computer vision applications, such as object detection, face recognition, and pedestrian detection.

Dong et al. [18] proposed a CNN-based SR algorithm, which directly learns an end-to-end mapping between the low- and high-resolution image pair. The three components of the pipeline in the recognition-based SR algorithms are represented as different layers

of CNN, which efficiently optimize the entire SR implementation through the CNN. The mapping is represented as a deep convolutional neural network (CNN) that takes the low-resolution image as the input and outputs the high-resolution one. The first step is patch extraction and representation. The recognition-based SR algorithms [17] use the densely extracted patches and then represent them by a set of pre-trained bases such as PCA, DCT, and Haar. This is equivalent of convolving the image by a set of filters, each of which is a basis. Thus, the first layer of the CNN can be expressed as:

$$F_1(Y) = \max(0, W_1 * Y + B_1), \tag{8}$$

where W_1 and B_1 represent the filter weights and biases, respectively, '$*$' denotes the convolution operation. W_1 is of size $c \times f_1 \times f_1 \times n_1$, corresponds to n_1 filters of spatial size $f_1 \times f_1$ and c stands for the number of channels in the image, that applies n_1 convolutions on the image. The output is composed of n_1 feature maps. B_1 is an n_1-dimensional bias vector, whose each element is associated with a filter. The second component of the recognition-based SR algorithm pipeline can be represented using the non-linear mapping step of CNN. As shown in Equation (8), the first layer extracts an n_1-dimensional feature vector for each patch. In the second operation, each of these n_1-dimensional vectors is mapped into an n_2-dimensional vetor. The operation of the second layer can be represented as:

$$F_2(Y) = \max(0, W_2 * F_1(Y) + B_2), \tag{9}$$

here W_2 is of size $n_1 \times f_2 \times f_2 \times n_2$, corresponds to n_2 filters of spatial size $n_1 \times f_2 \times f_2$, and B_2 is n_2-dimensional bias. Each of the output n_2-dimensional vectors is a representation of a high-resolution patch that will be used for SR reconstruction. Finally, the reconstruction step in the recognition-based SR algorithm pipeline produces the final HR image by averaging the overlapping high-resolution patches. The averaging can be considered as a pre-defined filter on a set of feature maps, where each position is the flattened vector form of a high-resolution patch.

$$F(Y) = W_3 * F_2(Y) + B_3, \tag{10}$$

where W_3 is of size $n_2 \times f_3 \times f_3 \times c$, corresponds to c filters of a spatial size $n_2 \times f_3 \times f_3$, and B_3 is a c-dimensional bias vector. The values of the parameters n_1, n_2, n_3, f_1, f_2, and f_3 used in the experiments are detailed in the experimental results, Section 4.4.

The super-resolution pre-processing technique is used to obtain high-resolution representation of the low-resolution face images as shown in Figure 7. We can see that the performance of the CNN-based super-resolution recovery method face is better than the performance of sparse-based super-resolution technique.

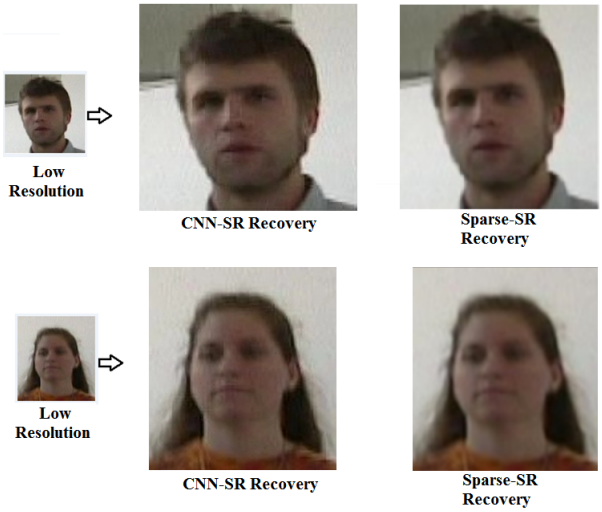

Figure 7. Super-resolution recovery of the LR face images.

3.2.2. Illumination and Pose Invariance

In this section, we explain the preprocessing steps for normalizing the low-resolution images with respect to illumination and pose variations.

It has been proven in the literature, that illumination variations are among the primary problems in biometric authentication. We adopted the Self-quotient image (SQI) [48] to normalize the illumination variations in the low-resolution facial images. SQI incorporates an edge-preserving filtering technique to minimize the spectral variations present in the illumination.

The Lambertian model can be factorized into two parts, the intrinsic part, and the extrinsic part:

$$I(x, y) = \rho(x, y)\, n(x, y)^T \cdot s = F(x, y) \cdot s, \tag{11}$$

where ρ is the albedo and n is the surface normals. $F = \rho n^T$ depends on the albedo and surface normal of an object and hence is an intrinsic factor, where F represents the identity of a face. However, s is the illumination and is an extrinsic factor. Separating the two factors and removing the extrinsic component is a key to achieve a robust face recognition by normalizing the effect of varying illumination.

The SQI image Q of an image I can be represented as:

$$Q = \frac{I}{\widehat{I}} = \frac{I}{P * I}, \tag{12}$$

where $\widehat{I}$ is the smoothed version of I, P is the smoothing kernel, and the division is pixel-wise. SQI [48] achieves the removal of extrinsic component s in Equation (11) through a two-step process. First, an illumination estimation step: the extrinsic factor is estimated to generate a synthesized smooth image, which has same illumination and shape as the input but a different albedo. Second, an illumination effect subtraction step: the illumination is normalized by computing the difference between the logarithms of the albedo maps of the input and the synthesized images, $(\log\rho_0 - \log\rho_1)$.

Pose variations present a major problem in real-world face recognition applications. Since the human face is approximately symmetric, if it is in the frontal pose with no rotations, the matrix containing the face image (F) will have the lowest rank. Employing the above-stated principle, Zhang et al. [49] proposed *transform invariant low-rank textures (TILT)* to normalize the pose of a rotated frontal face and remove minor occlusions.

TILT [49] tries to find a transformation (Euclidean, affine, or projective) matrix τ, through modeling the face rotation using an error matrix E, s.t. $\widehat{F} * \tau = F + E$, where $\widehat{F}$ represents the deformed and corrupted face and F is the corrected low-rank face image, by optimizing the following equation:

$$\min_{F,E,\tau} \text{rank}(F) + \gamma ||E||_o \quad \text{s.t.} \quad \widehat{F} * \tau = F + E \tag{13}$$

where $||E||_o$ is the l_0-norm of the error matrix, i.e., number of non-zero elements. It actually finds the corrected low-rank face image (F) with the lowest possible rank and the error with the lowest number of non-zero elements, which satisfy the above condition. γ trades off the rank of the matrix and the sparsity of the error.

Optimizing the rank function and the l_0-norm in the above equation is very challenging. Therefore, they are substituted by their convex surrogates. Since the rank of a matrix is equivalent to the number of its non-zero singular values, we can substitute the rank(F) by its nuclear norm $||F||_*$, which is the sum of its singular values. Moreover, l_0-norm is substituted by l_1-norm, which is the sum of the absolute values of the elements of the matrix. Additionally, the constraint $\widehat{F} * \tau = F + E$ is non-linear. By linearizing the constraint around its current estimate through an iterative process, the optimization problem becomes as follows:

$$\min_{F,E,\Delta\tau} ||F||_* + \gamma ||E||_1 \quad \text{s.t.} \quad \widehat{F} * \tau + \nabla\widehat{F}\Delta\tau = F + E, \tag{14}$$

where ∇ represents the Jacobian. Finally, we train a binary classifier using local features (Local Binary Pattern) to remove the false positive frames detected by the Adaboost face detector.

3.2.3. Registration and Synthesizing Low-Resolution Face Images

In this section, we describe the image registration of the pre-processed and normalized face regions, and synthesizing them using Curvelet and Inverse Curvelet transformation. We adopted the subspace-based holistic registration (SHR) method [50], which was proposed to perform registration on low-resolution face images. The majority of the automatic landmark-based registration methods can only perform accurate registration on high resolution images. However, SHR is able to obtain a user independent face model using Procrustes transformation by incorporating the image edges as feature vectors to register low-resolution face images. The best registration parameters are iteratively obtained through the downhill simplex optimization technique by maximizing the similarity score between the probe and the gallery image. The registration similarity is calculated using the probability that the probe and gallery face images are correctly aligned in a face subspace by computing the residual error in the dimensions perpendicular to the face subspace.

The first step of obtaining the subject independent face model to perform the registration is to compute the edges in the low-resolution facial image. Gaussian kernel derivatives of the LR face images are calculated in the x and y directions, respectively, using G_x and G_y as follows:

$$\begin{aligned} G_x(x,y) &= \frac{-x}{2\pi\sigma^4}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \\ G_y(x,y) &= \frac{-y}{2\pi\sigma^4}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \end{aligned} \tag{15}$$

The derivatives H_x and H_y of the images are obtained by convoluting the LR face image with G_x and G_y resulting in the "edge images" used for the registration purpose. Procrustes transformation is used to align the probe image to the gallery image by correcting the variations of scale by a factor f, rotatation with an angle α, and translation of $\mathbf{u}$, while preserving the distance ratios. Given a pixel location $\mathbf{p} = (x, y)^T$, the transformation $U_\vartheta \mathbf{p}$ on a pixel location can be represented as:

$$U_\vartheta \mathbf{p} = fR(\alpha)\mathbf{p} + \mathbf{u}, \tag{16}$$

where $\vartheta = \{\mathbf{u}, \alpha, f\}$ represent the registration parameters, and $R(\alpha)$ is the rotation matrix. The transformation of the entire probe image to perform the registration operation is obtained by applying U_ϑ on the computed "edge images" as follows:

$$T_\vartheta H(\mathbf{p}) = H(U_\vartheta^{-1}\mathbf{p}). \tag{17}$$

where $H = \sqrt{H_x^2 + H_y^2}$. Thus, a registered and aligned image, $T_\vartheta H(\mathbf{p})$, is obtained through backward mapping and interpolation by utilizing the optimal registration parameter ϑ found using simplex optimization technique.

To enhance the spectral features for face recognition, image synthesizing methods [51] are very popular in the literature. The synthesizing methods available in the literature can be broadly categorized into two classes, one performs the synthesis in the spatial domain and the other in the frequency domain. In this paper, we adopted the Curvelet-based image synthesis [52] which uses the Curvelet coefficients [53] to represent the face.

Curvelet transform has improved directional capability, better ability to represent edges, and other singularities along curves as compared to other traditional multiscale transforms, e.g., wavelet transform. First, curvelet transforms are applied to the sequence of registered face images. The smallest low-frequency components are represented by the coarse Curvelet coefficients and the largest high-frequency components are represented by

the fine Curvelet coefficients. For the image sequence $I_1, I_2, \ldots, I_n$, the Curvelet coefficients can be represented as $C_{I_i}\{j\}\{l\}$, where $i = 1, 2, \ldots, n$ represent the image sequence to be synthesized, and j, l are the scale and direction parameters, respectively. The components of the first scale where $j = 1$ represent the low-frequency parts of the face images, and the components associated to other scales ($j > 1$) represent the high-frequency parts. The minimum components between each $C_{I_i}\{1\}\{l\}$, where scale $j = 1$, and ($i = 1, 2, \ldots, n$), and the maximum components between each $C_{I_i}\{j\}\{l\}$, where ($j = 2, \ldots, 5$), and $i = 1, 2, \ldots, n$ are retained for the synthesized Curvet coefficients. Inverse Curvelet transformation of the synthesized Curvelet feature vector generates the synthesized image used for feature extraction.

3.2.4. Feature Extraction

We obtain LBP and Gabor features from the fused image and compare their performance for recognition. In the subsequent sections, we describe the LBP and Gabor feature extraction techniques.

The original LBP operator, introduced by Ojala et al. [54], is a powerful method for texture description. The operator labels the pixels of an image by thresholding the 3×3-neighborhood of each pixel with the center value and considering the result as a binary number. Then, the histogram of the labels can be used as a texture descriptor. See Figure 8 for an illustration of the basic LBP operator.

Later, the operator was extended to use neighborhoods of different sizes. Using circular neighborhoods and bilinearly interpolating the pixel values allow any radius and number of pixels in the neighborhood. For neighborhoods, we use the notation (P, R) which means P sampling points on a circle of radius of R. Figure 9 shows an example of the circular neighborhood (8,2). Another extension to the original operator uses what is called uniform patterns. A Local Binary Pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For example, 00000000, 00011110, and 10000011 are uniform patterns. Ojala et al. [54] noticed that in their experiments with texture images, uniform patterns account for a bit less than 90% of all patterns when using the (8,1) neighborhood and for around 70% in the (16,2) neighborhood.

An extension of LBP-based face description method is proposed by Ahonen et al. [55]. The facial image is divided into local regions ($k \times k$ window) and LBP texture descriptors are extracted from each region independently. The descriptors are then concatenated to form a global description of the face that describes the facial image in a high dimensional feature space. Window sizes used for experiment purposes are $k = 3, 5, 7$.

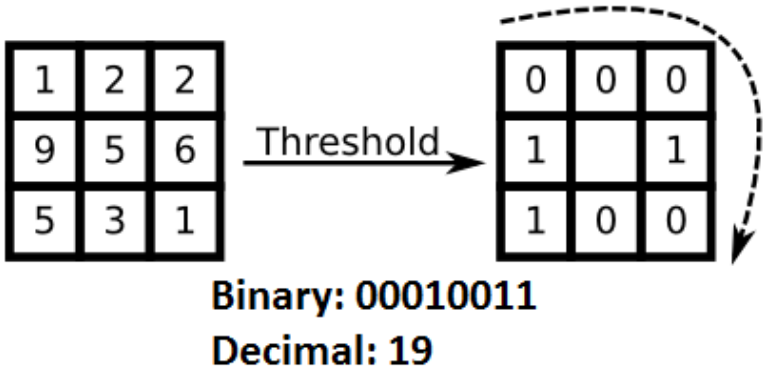

Figure 8. LBP feature, circular (8,1) neighborhood.

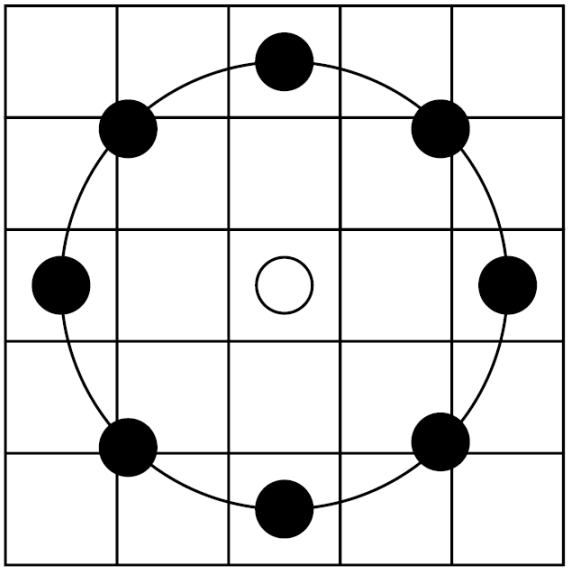

Figure 9. Circular (8,2) neigborhood.

2D Gabor filters [56] are used in a broad range of applications [57] to extract scale and rotation invariant feature vectors. In our feature extraction step, uniform down-sampled Gabor wavelets are computed for the detected regions using Equation (18), as proposed in [58]:

$$\psi_{\mu,\nu}(z) = \frac{||k_{\mu,\nu}||^2}{s^2} e^{\left(\frac{-||k_{\mu,\nu}||^2||z||^2}{2s^2}\right)} [e^{ik_{\mu,\nu}z} - e^{\frac{-s^2}{2}}], \tag{18}$$

where $z = (x, y)$ represents each pixel in the 2D image, $k_{\mu,\nu}$ is the wave vector, which can be defined as $k_{\mu,\nu} = k_\nu e^{i\varphi_\mu}$, $k_\nu = \frac{k_{max}}{f^\nu}$, k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain, $\varphi_\mu = \frac{\pi\mu}{2}$, and the value of s determines the ratio of the Gaussian window width to wavelength. Using Equation (18), Gabor kernels are generated from one filter using different scaling and rotation factors. In this paper, we used five scales, $\nu \in 0, ..., 4$ and eight orientations $\mu \in 0, ..., 7$. The other parameter values used are $s = 2\pi$, $k_{max} = \frac{\pi}{2}$, and $f = \sqrt{2}$.

Gabor features are computed by convolving each Gabor wavelet with the synthesized super-resolution pre-processed LR face images, as follows:

$$C_{\mu,\nu}(z) = T(z) * \psi_{\mu,\nu}(z), \tag{19}$$

where $T(z)$ is the face image, and $z = (x, y)$ represents the pixel location. The feature vector is constructed out of $C_{\mu,\nu}$ by concatenating its rows.

4. Experimental Results

In this section, we first describe the Face and Ocular Challenge Series (FOCS) [59] dataset. Then, we demonstrate the experiments and results of the frontal gait recognition and low-resolution face recognition followed by the score level fusion to obtain the multimodal recognition. While evaluating the performance of the biometrics recognition algorithm, we query each test data instance against all the subjects present in the gallery data, which results in the classification probabilities over all the subjects. However, while reporting Rank-1 accuracy, we only count the instances as true positive when the subject with the highest classification probability is the exact match with the test subject, thus Rank-1 accuracy is the precise measure which penalizes all strictly incorrect classification results, while a rank-k for $k > 1$ allows for some error.

4.1. FOCS Dataset

The video challenge dataset, Face and Ocular Challenge Series (FOCS) [59], contains video sequences of individuals, acquired on different days. Students from The University of Texas, Dallas, between the age group of 18 and 25, volunteered for the data collection. The FOCS dataset is collected in two sessions, where in the second duplicate session of data collection the subjects have a different hairstyle, different clothing, and may be otherwise different in appearance.

The FOCS database contains a variety of still images and videos of a large number of individuals taken in a variety of contexts. For our experiments, we used the frontal walking video sequences. The videos in the FOCS database were collected using a Canon Optura Pi digital video camera, this applies a single progressive scan CCD digitizer resulting in minimal motion aliasing artifacts. The videos were stored in DV Stream format at 720×480 with 24-bit color resolution and 29.97 frames per second. In the frontal walking video sequences, the subject walks parallel to the line of sight of the camera, approaching the camera, but veering off to the left while reaching in front of the camera. These frontal walking video sequences capture the subject from the start point until he/she goes out of view. Thus, the time varies somewhat for each subject due to their walking speed, but on average it is approximately 10 seconds. The FOCS frontal walking video sequences contain videos that are acquired from 136 unique subjects. The number of samples per subject varies. Out of 136 subjects, 123 subjects have at least 2 videos. We used data from these 123 subjects for our experiments, where one of the video clips is randomly chosen for training and the other is used for testing.

4.2. Experimental Setup

To perform the multimodal recognition, we first segment the frontal gait silhouette from the background and detect the low-resolution face images from the frontal walking video sequences as described in Section 3. In order to evaluate the proposed algorithm, we perform the frontal gait recognition and the low-resolution face recognition experiments as two separate components. Later, we use the match score level fusion scheme to fuse the individual recognition results.

4.2.1. Frontal Gait Recognition

Once the binary gait silhouette is acquired, we obtained the scale and translation invariant 3D moments of the spatio-temporal GEI or the periodic gait volume, and the average number of frames in the identified gait cycles in the frontal walking video clips to prepare a high dimensional feature vector as described in Section 3.1.1. The frontal gait features were classified using a k-nearest neighbor classifier (k-NN), where a test gait feature vector belongs to the class that minimizes the similarity distance between the gallery and the probe gait feature vector.

We performed quantitative experiments using different moment orders $O = p_1 \cdot p_2 \cdot p_3$. The best recognition performance was obtained when $p_1 = p_2 = p_3 = 10$. The results of frontal gait recognition are presented in Table 1. We can see that the recognition performance when using the 3D moments of periodic gait volume is better than the performance when using the average movement speed feature representation. However, concatenating the feature vectors together improved the gait recognition performance. Table 2 shows a comparison between the frontal gait recognition performance achieved by the proposed approach and the recognition accuracy of the state-of-the techniques on the FOCS dataset. The proposed frontal gait recognition system achieves a rank-one recognition rate of 93.5% on the 123 subjects of FOCS dataset.

Table 1. Frontal gait recognition.

Feature Vectors Used	Rank-1 Accuracy
3D Moments	88.62% (109 out of 123)
Average movement speed	69.11% (85 out of 123)
3D Moments and Average movement speed	**93.5%** (115 out of 123)

Table 2. Comparison of frontal gait recognition accuracy.

Method	Rank-1 Frontal Gait Recognition Accuracy
Wang et al. [3]	69.11%
Chen et al. [4]	89.43%
Goffredo et al. [26]	91.06%
This Work	**93.50%**

4.2.2. Low-Resolution Face Recognition

The first step of LR face recognition is to detect the low-resolution faces using the Adaboost detector from the video surveillance frames as described in Section 3. The proposed LR face recognition Algorithm 1 is described in Section 3.2. After employing the CNN-based super-resolution technique to obtain the high-resolution equivalent of the LR faces, we perform the illumination and pose normalization steps. The sizes of the pre-processed face images vary between 40×40 pixels and 180×180 pixels. To effectively leverage the high-frequency information present in the pre-processed face images, we separate the face images into two classes. The face images of size less than 96×96 pixels are labeled as Class – 1 and those that are greater than 96×96 pixels are labeled as Class – 2. We use the face image of size 72×72 pixel in Class – 1 as the base or template image to apply the SHR registration technique [50], described in Section 3.2.3, to register all the face images that belong to Class – 1 after rescaling them to 72×72 pixels. Similarly, the face image of size 120×120 pixels in Class – 2 is used as the base or template image to apply the SHR registration technique [50] to register all the face images that belong to Class – 2 after rescaling them to 120×120 pixel. After performing the image synthesis using the Curvelet coefficients as described in Section 3.2.3 of the face images in Class – 1 and Class – 2 separately, we obtain two synthesized face images for each surveillance video clip. We extract the LBP and Gabor feature vectors, as mentioned in Section 3.2.4, from the two synthesized face images and perform feature concatenation to obtain the composed LBP and Gabor features, which represent the LR face in the surveillance video clip. The obtained LBP and Gabor feature vectors are used separately to compare their performance in LR face recognition. For each of the 123 subjects used in the performance evaluation, the feature vector obtained from one randomly chosen surveillance video clip is used to build the model and the one obtained from the other video is used for testing.

Algorithm 1 Low-Resolution Face Recognition

1. Detect faces in the video surveillance frames.
2. Use a Super-resolution technique to obtain High-resolution from the Low-resolution detected face images.
3. Perform illumination and pose normalization.
4. Register the pre-processed and normalized face regions, followed by synthesizing them using Curvelet and Inverse Curvelet transformations.
5. Extract Local Binary Pattern (LBP) and Gabor features from the synthesized image.
6. Perform face recognition using the extracted features.

We compare the performance of the proposed LR face recognition technique using the CNN-based super-resolution [18] with the following baseline algorithms. It is worth noting that all the comparisons are based on the same training/test set.

(1) LR face recognition without any super-resolution pre-processing technique.
(2) LR face recognition using Bicubic Interploation super-resolution pre-processing technique.
(3) LR face recognition using Sparse super-resolution [17] pre-processing technique.

The obtained LR face features were classified using a k-nearest neighbor classifier (k-NN), where a test LR facial feature vector belongs to the class that minimizes the similarity distance between the gallery and the probe feature vector. The result of low-resolution face recognition is presented in Table 3. We can see that the performance when using the local feature representation LBP is better than the performance when using global feature representation or Gabor features. Moreover, by employing the CNN-based super-resolution technique the LR face recognition performance is increased to 82.91% compared with 72.36% without any SR pre-processing of the LR face images.

Table 3. Low-resolution face recognition.

Features Used	Super Resolution Technique	Rank-1 Accuracy
LBP	None	72.36% (89 out of 123)
Gabor	None	70.73% (87 out of 123)
LBP	Bicubic	73.98% (91 out of 123)
Gabor	Bicubic	71.54% (88 out of 123)
LBP	Sparse	75.61% (93 out of 123)
Gabor	Sparse	72.36% (89 out of 123)
LBP	SRCNN	**82.92%** (102 out of 123)
Gabor	SRCNN	79.67% (98 out of 123)

4.3. Multimodal Recognition Accuracy

Score level fusion techniques are very popular in multimodal biometrics applications specifically in the application of fusing face and gait [2,27]. In our experiment, results from the different classifiers were combined directly using the Sum, Max, and Product rules.

To prepare for fusion, the matching scores obtained from the different matchers are transformed into a common domain using a score normalization technique. Later, the score fusion methods are applied. We have adopted the Tanh score normalization technique [60], which is both robust and efficient, defined as follows:

$$s_j^n = \frac{1}{2}\left\{\tanh(0.01(\frac{s_j - \mu_{GH}}{\sigma_{GH}})) + 1\right\}, \tag{20}$$

where s_j and s_j^n are the match scores before and after normalization, respectively. μ_{GH} and σ_{GH} are the mean and standard deviation estimates of the actual score distribution given by Hampel estimators [61], respectively. Hampel's estimators are based on the influence functions ψ which are odd functions and can be defined for any x (matching score, s_j, in this paper) as follows:

$$\psi(x) = \begin{cases} x, & 0 \leq |x| < a, \\ a\ \mathrm{sgn}(x), & a \leq |x| \leq b, \\ \frac{a(r-|x|)}{r-b}\ \mathrm{sgn}(x), & b \leq |x| \leq r, \\ 0, & r \leq |x|, \end{cases} \tag{21}$$

where

$$\mathrm{sgn}(x) = \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{otherwise}, \end{cases} \tag{22}$$

In Equation (21), the values of a, b, and r in ψ reduce the influence of the scores at the tails of the distribution during the estimation of the location and scale parameters, i.e., μ_{GH} and σ_{GH} in Equation (23). The normalized match scores of synthesized face images of the

gallery and probe and the normalized match scores of gaits of the gallery and probe from the same video clips are fused based on different match score fusion techniques. Let s^n_{jF} and s^n_{jG} be the normalized match scores obtained from a specific video clip for the face and gait, respectively. The unknown test subject is classified to class C if the fused match score corresponding to the class C is maximum compared to all other classes in the gallery:

$$FR\{s^n_{CF}, s^n_{CG}\} = \max FR\{s^n_{jF}, s^n_{jG}\}; j \in (1, 2, \ldots, N) \quad (23)$$

where FR{,} represents the fusion rule, and N represents the number of enrolled individuals in the gallery. In this paper, we use Sum, Max, and Product rules.

The results of the fused multimodal recognition are presented in Table 4. We can see that the fusion based on the Sum rule of the frontal gait and the LR face results in the best recognition accuracy.

Table 4. Comparison of multimodal recognition fusion scheme.

Fusion Rule	Rank-1 Accuracy
Sum Rule	**95.9%** (118 out of 123)
Max Rule	94.3% (116 out of 123)
Product Rule	93.5% (115 out of 123)

4.4. Parameter Selection for the CNN Super Resolution

We tested the performance of the proposed LR face recognition with different parameters of the convolution neural network. The number of layers in the CNN network is varied between 3 and 5, where the best performance was obtained when using 3 layer architecture. The recognition-based super-resolution algorithm has three distinct steps, which signifies the optimal performance of the CNN with 3 layers. Experiments are conducted by varying the numbers of filters n_1 and n_2 (refer to Equations (9) and (10)) of the CNN architecture. Three sets of network parameters were used for experimental purposes (n_1 = 32 and n_2 = 16), (n_1 = 64 and n_2 = 32), and (n_1 = 128 and n_2 = 64). The best performance was achieved with the parameters (n_1 = 128 and n_2 = 64). The Super resolution restoration speed decreases with the increase of the size of the filters. To obtain a reasonable trade off we set the number of the filters n_1 and n_2 to 128 and 64, respectively. Moreover, the size of filters f_1, f_2, and f_3 (refer to Equations (8)–(10)) are varied between (9, 1, 5), (9, 3, 5), and (9, 5, 5). The best accuracy and performance trade off was obtained using the parameter values of f_1 = 9, f_2 = 3, and f_3 = 5. With the above-mentioned parameter settings, 8×10^8 iterations of backpropagations were needed to achieve convergence.

5. Conclusions

We proposed a system for highly accurate multimodal human identification from low-resolution video surveillance footage through LR face and frontal gait recognition using a single biometric data source, i.e., frontal walking surveillance video. Using the trained Adaboost detector, we automatically detect the LR face images. The frontal gait binary silhouette's are segmented using the fast object segmentation algorithm. We proposed an approach for accurate identification of the gait cycles in the entire gait video clip using only frontal gait information, then we extract the average movement speed and the shape feature. The detected LR face images are pre-processed using super-resolution techniques to obtain the high-resolution representation. This is followed by illumination and pose normalization, and image synthesis through registration. Finally, Gabor and LBP features are extracted from the synthesized face images. The nearest neighbor classifier is used to obtain modality specific rank-1 recognition for each modality. Then, the individual recognition results are fused through the score level fusion. The results indicate that combining the LR face and the frontal gait modalities produce the best recognition Rank-1 accuracy compared to the performance of each modality.

Author Contributions: Conceptualization, S.M., M.A.-M. and S.S.A.; Methodology Implementation, S.M., Manuscript Drafting, S.M., M.A.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Shakhnarovich, G.; Darrell, T. On probabilistic combination of face and gait cues for identification. In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 21–21 May 2002; pp. 169–174.
2. Kale, A.; Roychowdhury, 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 '04), Montreal, QC, Canada, 17–21 May 2004; Volume 5, p. V–901–4.
3. Wang, L.; Tan, T.; Hu, W.; Ning, H. Automatic gait recognition based on statistical shape analysis. *IEEE Trans. Image Process.* **2003**, *12*, 1120–1131. [CrossRef]
4. Chen, S.; Gao, Y. An Invariant Appearance Model for Gait Recognition. In Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, Beijing International Convention Center, Beijing, China, 2–5 July 2007; pp. 1375–1378.
5. Bronstein, A.; Bronstein, M.; Kimmel, R. Three-Dimensional Face Recognition. *Int. J. Comput. Vis.* **2005**, *64*, 5–30. [CrossRef]
6. Etemad, K.; Chellappa, R. Discriminant analysis for recognition of human face images. In *Audio- and Video-Based Biometric Person Authentication*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 1997; pp. 125–142.
7. Lu, C.; Tang, X. Surpassing Human-Level Face Verification Performance on LFW with GaussianFace. *arXiv* **2014**, arXiv:1404.3840.
8. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1701–1708.
9. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef]
10. Wang, Z.; Miao, Z.; Jonathan Wu, Q.M.; Wan, Y.; Tang, Z. Low-resolution face recognition: A review. *Vis. Comput.* **2014**, *30*, 359–386. [CrossRef]
11. Wilman, W.W.Z.; Yuen, P.C. Very low resolution face recognition problem. In Proceedings of the 2010 Fourth IEEE International Conference on Biometrics, Theory Applications and Systems (BTAS), Washington, DC, USA, 27–29 September 2010; pp. 1–6. [CrossRef]
12. Sarkar, S.; Phillips, P.J.; Liu, Z.; Vega, I.R.; Grother, P.; Bowyer, K.W. The humanID gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 162–177. [CrossRef]
13. Man, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 316–322.
14. Kusakunniran, W.; Wu, Q.; Li, H.; Zhang, J. Multiple views gait recognition using View Transformation Model based on optimized Gait Energy Image. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), Kyoto, Japan, 27 September–4 October 2009; pp. 1058–1064.
15. Zheng, S.; Zhang, J.; Huang, K.; He, R.; Tan, T. Robust view transformation model for gait recognition. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2073–2076.
16. Ren, C.X.; Dai, D.Q.; Yan, H. Coupled Kernel Embedding for Low-Resolution Face Image Recognition. *IEEE Trans. Image Process.* **2012**, *21*, 3770–3783. [PubMed]
17. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image Super-Resolution Via Sparse Representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef] [PubMed]
18. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef]
19. Shekhar, S.; Patel, V.M.; Chellappa, R. Synthesis-based recognition of low resolution faces. In Proceedings of the 2011 International Joint Conference on Biometrics (IJCB), Washington, DC, USA, 11–13 October 2011; pp. 1–6.
20. Yu, J.; Bhanu, B. Super-resolution Restoration of Facial Images in Video. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 4, pp. 342–345.
21. Jia, K.; Gong, S. Generalized Face Super-Resolution. *IEEE Trans. Image Process.* **2008**, *17*, 873–886.
22. Lee, T.K.M.; Belkhatir, M.; Sanei, S. A comprehensive review of past and present vision-based techniques for gait recognition. *Multimed. Tools Appl.* **2014**, *72*, 2833–2869. [CrossRef]
23. Lee, L.; Grimson, W.E.L. Gait analysis for recognition and classification. In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 21 May 2002; pp. 148–155.
24. 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]
25. Davis, J.W.; Bobick, A.F. The representation and recognition of human movement using temporal templates. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; pp. 928–934.

26. Goffredo, M.; Carter, J.N.; Nixon, M.S. Front-view Gait Recognition. In Proceedings of the 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS 2008, Washington, DC, USA, 29 September–1 October 2008; pp. 1–6.
27. Lu, H.; Wang, J.; Plataniotis, K.N. Chapter A Review on Face and Gait Recognition: System, Data and Algorithms. In *Advanced Signal Processing: Theory and Implementation for Sonar, Radar, and Non-Invasive Medical Diagnostic Systems*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2009; pp. 303–325.
28. Zhou, X.; Bhanu, B. Feature Fusion of Face and Gait for Human Recognition at a Distance in Video. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 4, pp. 529–532.
29. Xin, G.; Kate, S.-M.; Liang, W.; Ming, L.; Qiang, W. Context-aware fusion: A case study on fusion of gait and face for human identification in video. *Pattern Recognit.* **2010**, *43*, 3660–3673.
30. Papazoglou, A.; Ferrari, V. Fast Object Segmentation in Unconstrained Video. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1777–1784.
31. Brox, T.; Malik, J. Chapter Object Segmentation by Long Term Analysis of Point Trajectories. In Proceedings of the 11th European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 282–295.
32. Sundaram, N.; Brox, T.; Keutzer, K. *Dense Point Trajectories by GPU-Accelerated Large Displacement Optical Flow*; Technical Report UCB/EECS-2010-104; EECS Department, University of California: Berkeley, CA, USA, 2010.
33. Rother, C.; Kolmogorov, V.; Blake, A. GrabCut -Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314. [CrossRef]
34. Lee, Y.J.; Kim, J.; Grauman, K. Key-segments for Video Object Segmentation. In *ICCV '11, Proceedings of the 2011 International Conference on Computer Vision*; IEEE Computer Society: Washington, DC, USA, 2011; pp. 1995–2002.
35. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 511–518.
36. Phillips, P.; Moon, H.; Rizvi, S.; Rauss, P. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1090–1104. [CrossRef]
37. Murray, M.P.; Drought, A.B.; Kory, R.C. Walking Patterns of Normal Men. *J. Bone Jt. Surg.* **1964**, *46*, 335–360. [CrossRef]
38. Perry, J.; Burnfield, J.M. *Gait Analysis: Normal and Pathological Function*; Slack Incorporated: West Deptford Township, NJ, USA, 2010.
39. Little, J.J.; Boyd, J.E. Recognizing People by Their Gait: The Shape of Motion. *Videre J. Comput. Vis. Res.* **1998**, *1*, 1–32.
40. Stephenson, J.L.; Serres, S.J.D.; Lamontagne, A. The effect of arm movements on the lower limb during gait after a stroke. *Gait Posture* **2010**, *31*, 109–115. [CrossRef] [PubMed]
41. Jackson, K.M.; Joseph, J.; Wyard, S.J. The upper limbs during human walking. part 2: Function. *Electromyogr. Clin. Neurophysiol.* **1983**, *23*, 435–446. [PubMed]
42. Lee, T.K.M.; Belkhatir, M.; Lee, P.A.; Sanei, S. Fronto-normal gait incorporating accurate practical looming compensation. In Proceedings of the 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
43. Flusser, J.; Suk, T. Pattern recognition by affine moment invariants. *Pattern Recognit.* **1993**, *26*, 167–174. [CrossRef]
44. Hu, M. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theory* **1962**, *8*, 179–187.
45. Maity, S.; Abdel-Mottaleb, M. 3D Ear Segmentation and Classification Through Indexing. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 423–435. [CrossRef]
46. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
47. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
48. Wang, H.; Li, S.Z.; Wang, Y.; Zhang, J. Self quotient image for face recognition. In *ICIP '04, Proceedings of the 2004 International Conference on Image Processing*, Singapore, 24–27 October 2004; Volume 2, pp. 1397–1400.
49. Zhang, Z.; Ganesh, A.; Liang, X.; Ma, Y. TILT: Transform Invariant Low-Rank Textures. *Int. J. Comput. Vis.* **2012**, *99*, 1–24. [CrossRef]
50. Boom, B.J.; Speeuwers, L.J.; Veldhuis, R.N.J. Subspace-Based Holistic Registration for Low-Resolution Facial Images. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 14. [CrossRef]
51. Mitchell, H.B. *Image Fusion: Theories, Techniques and Applications*; Springer: Berlin, Germany, 2010.
52. Xu, X.; Liu, W.Q.; Li, L. Low Resolution Face Recognition in Surveillance Systems. *J. Comput. Commun.* **2014**, *2*, 70–77.
53. Candès, E.; Demanet, L.; Donoho, D.; Ying, L. Fast Discrete Curvelet Transforms. *Multiscale Model. Simul.* **2006**, *5*, 861–899. [CrossRef]
54. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
55. Ahonen, T.; Hadid, A.; Pietikainen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef]
56. Liu, C.; Wechsler, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.* **2002**, *11*, 467–476. [PubMed]

57. Urolagin, S.; Prema, K.V.; Subba Reddy, N. Rotation invariant object recognition using Gabor filters. In Proceedings of the International Conference on Industrial and Information Systems, 2010s of the International Conference on Industrial and Information Systems, Mangalore, India, 29 July–1 August 2010; pp. 404–407.
58. Yang, M.; Zhang, L. Gabor Feature Based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In *Proceedings of the European Conference on Computer Vision*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2010; Volume 6316, pp. 448–461.
59. O'Toole, A.J.; Harms, J.; Snow, S.L.; Hurst, D.R.; Pappas, M.R.; Ayyad, J.H.; Abdi, H. A video database of moving faces and people. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 812–816. [CrossRef] [PubMed]
60. Ross, A.A.; Nandakumar, K.; Jain, A.K. *Handbook of Multibiometrics*; Springer: Berlin, Germany, 2006.
61. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. *Robust Statistics: The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*; Wiley-Interscience, John Wiley and Sons Inc.: Hoboken, NJ, USA, 2005.