Chapter 1

Introduction

1.1 Person Re-Identification

Person re-identification refers to the process of finding one-one correspondences among person images captured using multiple cameras having overlapping/non-overlapping fields of view. Figure 1.1explains the concept of person re-identification. It shows images of the same subject captured at different times using two different cameras. It can be seen from the figure that the pose of the person varies in the two images, and hence standard shape matching techniques or non-invasive biometric recognition techniques

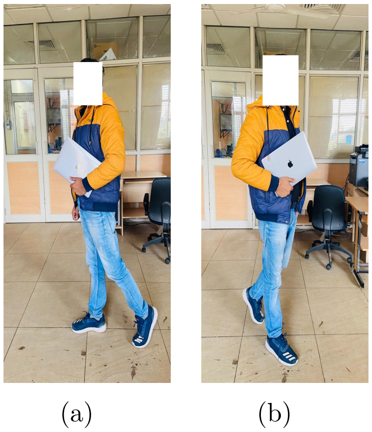

Figure 1.1: Images of same person captured using two different cameras at different times

like gait recognition cannot be directly used for re-identification. Also, the lighting conditions may be different in the locations where the cameras are installed, or there may be other static and/or dynamic objects that occlude the target subject. Intelligent surveillance applications often require tracking a subject across multiple cameras in a non-invasive and automated manner as he/she moves through the surveillance zone. Developing a suitable Computer Vision-based person re-identification approach using the latest learning techniques can potentially accomplish the above objective and also help in providing improved public safety. To date, several interesting approaches to person re-identification have been developed by researchers worldwide, most of which work using RGB images/videos. However, the practical application of Computer Vision-based re-identification techniques in surveillance sites is still limited. In this thesis, we aim to develop effective algorithms with a focus on improving the state-of-the-art, thereby making Computer Vision-based re-identification more suitable for application in real-life surveillance scenarios.

1.1.1 Possible Scenarios and Application Areas for Person Re-Identification

Depending on the application scenario, person re-identification approaches can be categorized into either of the following categories: (i) closed-set and open-set, or (ii) long-term and short-term.

Closed Set and Open Set Re-Identification Scenario: In a closed-set scenario, the gallery set remains constant, i.e., the gallery set does not evolve with time. Re-identification has to be done by comparing a target subject against a gallery of a fixed number of subjects. These kinds of re-identification scenarios can be seen at the exit points of stadiums and movie halls where several people assemble in an area and leave the place one by one through the exit gates. In an open-set re-identification scenario, the gallery set evolves with time, i.e., the number of subjects to re-identify keeps on changing with time. It is more challenging compared to closed-set re-identification since in this scenario a person may enter or leave the monitoring zone at any time. In case a new subject enters, the gallery set must be expanded to store the features corresponding to the new subject. Similarly, if a person leaves the monitoring zone, the gallery set can be reduced by eliminating the corresponding features. These kinds of re-identification scenarios can be seen at surveillance sites such as airports, railway stations, and shopping malls where several people enter and leave the zone continuously.

Short-term and Long-term Re-identification: Short-term re-identification refers to scenarios where subjects' clothing conditions and, in general, appearances do not change in the images captured by the different cameras, as shown in Figure 1.1. Example of suitable application sites for short-term re-identification are call-centers or office where a subject's clothing remains the same throughout the day. Such approaches focus more on extracting appearance-based descriptors. On the other hand, long-term re-identification does not depend on a subject's appearance, i.e., clothing conditions and it can be used to perform re-identification even if there is a change in the color-based appearance of the subject in the two camera views. This is typically used if the time between the capturing of the first and second set of images is long. These kinds of re-identification systems focus either on capturing shape features and/or other biometric features like gait and face. This thesis focuses on short-term re-identification scenario and provide plausible solutions to some important challenges associated with this problem.

1.1.2 Need for Development of Automated Re-id Technique

Manual tracking of persons in surveillance sites is not a suitable option if the order of individuals entering and leaving the fields of view of the different cameras is different. Also, surveillance sites like railway stations and airports usually consist of a large network of surveillance cameras that continuously capture high volumes of image/video data. Often it becomes quite inconvenient to manually monitor these large volumes of data and derive meaningful information from the camera-captured images mostly due to human fatigue. Manual inspection is also cost-intensive, error-prone, and non-reliable. This demands the development of automated surveillance solutions so that the image/video captured by the cameras can be analyzed more accurately and robustly in a cost-effective manner. Although several automated techniques for re-identification have been developed to date, to the best of our knowledge, no such approaches can be used potentially in practical/ real-life scenarios due to their limitations in effective handling of real-life situations like occlusion, illumination variation, etc. In this thesis, we have made an attempt to improve the state-of-the-art by developing automated re-identification techniques that can be conveniently applied in surveillance sites to track individuals as they move through the surveillance zone monitored by multiple cameras. Since Deep Learning-based algorithms are known to be robust and accurate in handling various image analysis tasks, in this thesis also the solutions proposed for the different re-identification challenges employ the powerful generalization capability of Deep Neural Networks.

1.1.3 Data Acquisition Devices to Carry-Out Re-id in Surveillance Applications

The image/video data in surveillance sites can be captured using various types of cameras depending upon the application scenario. A few popularly used camera types are explained next.

Thermal Camera: Thermal cameras capture infrared energy and use the data to create images through digital or analog outputs. Thermal cameras pick up the heat emitted from the objects to construct an infra-red image of the scene and can be suitably used to carry out surveillance at night even in the absence of any other light source. Although both heat and light are parts of the electromagnetic spectrum, a camera that can detect visible light cannot capture thermal energy and vice versa. Transair

Thermal scanner and camera, Fluke Compact PTi120, and Thermal Imaging Camera Uni-T UT-165A are a few popular thermal camera models, and Samsara manufactures, Senspex, Inc., AEMC Instruments, and Deep Blue Integration, Inc. are some of the leading thermal camera manufacturers.

Depth Camera: A depth camera, on the other hand, forms images in which each pixel value corresponds to the depth of the corresponding object point from the camera. Some organizations manufacture cameras consisting of both RGB and depth-sensing systems, e.g., Microsoft Kinect. These are also popularly referred to as the RGB-D cameras and these provide a depth channel along with the red, green, and blue channels. Popular depth-sensing technologies include Light detection and ranging (LIDAR), Time-of-flight (ToF) [4]. The leading depth camera manufacturers are Airmar Technology, Bluetechnix Group GmbH, and Decagon Devices, while the popular depth camera models are Microsoft Kinect, Intel RealSense L515, and Intel RealSense D455.

RGB Camera: This type of camera is equipped with a standard CMOS sensor through which the colored images of persons and objects are acquired in three complementary color channels, namely, red, green, and blue channels. The resolution of the captured images is usually expressed in terms of megapixels. The leading manufacturers of RGB cameras are Nikon, Sony, Canon, and Olympus, and some of their popular camera models are Sony DSC-H2, Olympus OM-D E-M1 Mark II, Nikon Z7, and Sony Alpha ILCE-QX1.

As already explained in Section 1.1.1, in short-term re-identification applications, color information plays an important role in constructing the descriptors for matching, and hence we consider RGB images only to derive the features for performing re-identification.

1.1.4 Challenges in Computer Vision-Based Person Re-id Applied to Surveillance Sites

The challenges associated with computer vision-based person re-identification can be broadly classified in two categories: (i) system-level challenges and (ii) component-level challenges.

System-level Challenges: The first and foremost challenge of re-identification, namely, the system-level challenge is establishing one-one correspondences among persons in the fields of view of the different cameras. As long as a subject stays in the monitoring zone of a particular camera, he/she can be easily tracked and his/her actions can be detected. However, once the subject moves away from the field of view of a camera and enters another, he/she must be identified properly from all the subjects captured by the previous camera. The challenge that needs to be addressed here is determining the true class of a target subject by matching images that are separated in both space and time and providing a consistent tracking label to the subject.

Component level Challenges: The second type of challenge is extracting accurate descriptors for matching the images and establishing the one-one correspondence. The same person may appear in the two camera views with two different poses. Also, the lighting conditions may vary in the locations where the two cameras are positioned which may cause a substantial change in the RGB appearance of the subject. There may be also the presence of occlusion due to other static and dynamic objects in between the person and the camera. Some of the major component-level challenges that a computer vision-based person re-identification algorithm has to tackle are listed next:

- variation in lighting conditions,
- varying poses of the same subject in the two frames captured by the two cameras,
- similar clothing of two different persons resulting in the extraction of similar appearance descriptors, and
- unavailability of complete person silhouette due to partial occlusion.

1.2 Motivation of the Work

The application of Deep Learning to solve the challenges associated with person re-identification (as discussed in the previous section) has already been studied by several researchers in the past and encouraging results have been obtained. We make an extensive literature survey of Computer Vision-based person re-identification using traditional and modern Deep Learning techniques and are also motivated to propose Deep Learning-based solutions to the various challenges associated with the problem of person re-identification. From the literature survey, we have identified some scopes for further research as explained next:

Developing Improved Deep Features for Re-identification: Most existing Deep Learning-based solutions to re-identification usually employ a Deep Network to compute the image features for re-identification. However, the use of dilation by increasing the receptive fields of the filters in a Deep Convolutional Network can capture detailed multi-scale features from images and are expected to be capture more effective low-level details from an input image, which may be useful for re-identification. Very recently, a few studies have also used multi-scaled features to perform re-identification [5,6]. Deep Neural Network-based re-identification with dilation in the convolutional layers may perform re-identification accurately that needs to be further studied.

Effective Combination of Passive and Modern Re-identification Approaches: Existing methods using Deep Learning are computationally time-intensive due to using multiple convolutional layers with a high number of trainable parameters. This may be overcome by following a hierarchical classification scheme that removes dissimilar elements after each level. For example, color-based features can be used for matching with the complete gallery set at an initial level and reduce the gallery set by retaining only the samples with similar appearance as that of the target. Next, re-identification can be carried out on the reduced gallery set using any suitable classifier. Such an approach will also prevent the prediction of the re-identification model to get biased towards a different but structurally similar object.

Fusing Information from Multiple Feature Extractors for Video-Based Person Re-identification: Existing video-based person re-identification methods usually employ a single model to extract spatio-temporal features for re-identification. Different time-series Neural Networks, such as RNN, LSTM, GRU, etc., are capable of capturing different aspects of the motion features even when trained on the same video sequence. Hence, a fusion of features from different time-series feature extractors may help in capturing better dynamic information of human motion. The effectiveness of using an ensemble of time-series Neural Networks in video-based person re-identification has not been studied in the past and it needs further attention.

Handling Occlusion in Re-identification: Any real-life surveillance application, such as tracking, re-identification, gait recognition, etc., suffers from the problem of occlusion caused by other static or dynamic objects. Occlusion can occur anywhere in a video sequence and prevents the camera from capturing clean silhouette of the target subject/s. Recently, GANs are being extensively used in various image reconstruction tasks [7–9], and have also been studied to reconstruct occluded frames in an input sequence for improved re-identification [10–12]. However, the existing solutions to person re-identification in the presence of occlusion suffers from low accuracy which limits their practical application. This is mostly due to the fact that these methods use a single Deep Network to perform occlusion reconstruction and re-identification simultaneously. It appears that a better accuracy may be achieved if two separate networks are used dedicated towards occlusion handling and re-identification which needs to be studied in the future. Also, to the best of our knowledge, for video-based person re-identification, there does not exist any method that reconstructs the occluded frames by exploiting the spatio-temporal information from the previous frames of the video. Reconstruction using the spatio-temporal content of sequential images frames is likely to be more effective compared to frame-specific spatial information-based reconstruction, as done in most of the previous studies, and this is also another major scope of the study. In this thesis, we have made attempts to provide effective Deep Learning-based solutions to address the above-mentioned scopes. The main contributions made in the thesis are highlighted next.

1.3 Contributions of the Thesis

In the thesis, we have discussed effective approaches to person re-identification both from sequential and non-sequential image frames. We have also considered the challenging situation of occlusion that can occur in any real-life surveillance site. Additionally, we have constructed a new video-based re-identification data set and made it publicly available to the research community along with all our trained models for further comparisons. The main contributions of the thesis chapter-wise are as follows:

1.3.1 Developing Improved Person Re-identification Approaches from Still Images

We consider a closed-set re-identification scenario in which a set of non-sequential frames of a person are available to perform re-identification. Figure 1.2 explains the scenario diagrammatically. With reference to the figure, we have a set of frames in varying poses corresponding to the gallery subjects *P1*, *P2*, *P3*. From each set, we compute spatial domain features and store in a database of gallery features. During testing, one among the gallery subjects appear in the field of view of a second camera and his/her correct class is predicted by extracting similar spatial domain features and comparing with the database of gallery features. We propose two effective Deep Learning-based approaches to person re-identification that can be potentially applied to scenarios similar to that discussed above. The highlights of the contribution are presented point-wise next:

- Development of effective features for re-identification using spatial domain features from image frames

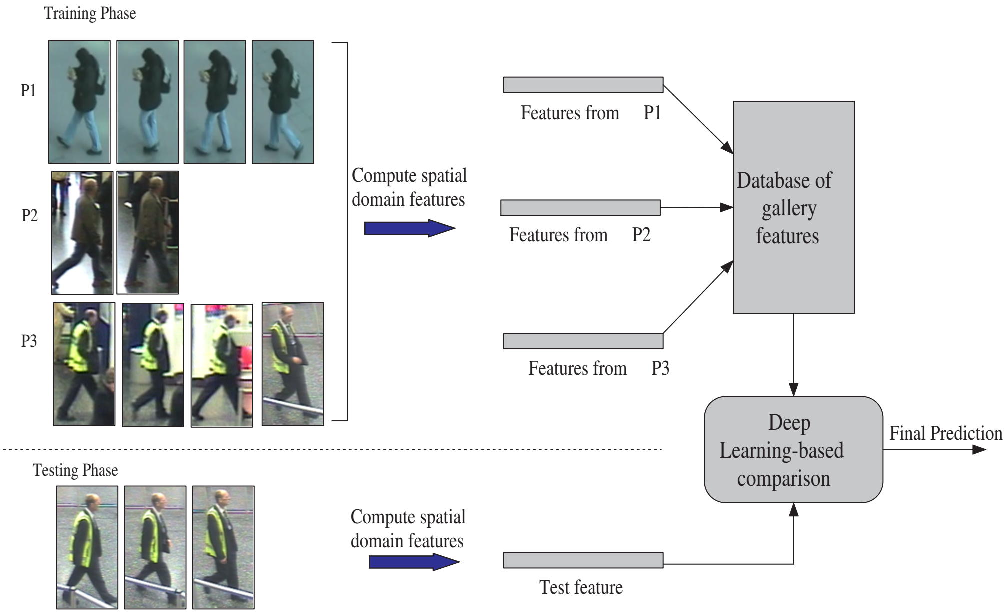
Training Phase
P1
P2
P3
Compute spatial domain features
Features from P1
Features from P2
Features from P3
Database of gallery features
Deep Learning-based comparison
Final Prediction
Testing Phase
Compute spatial domain features
Test feature

Figure 1.2: Scenario where spatial features extracted from non-sequential frames are used for re-identification

- Carrying out silhouette part-based analysis to improve the accuracy of prediction
- Performing experiments and thorough comparison with existing approaches to evaluate its effectiveness and constructing a database of 41 subjects and making it publicly available.

First, we develop a technique based on multi-scale feature extraction using dilation to capture the low-level information from input images properly. Here, we carry out silhouette part-based feature extraction by dividing a silhouette into multiple sub-parts and constructing a feature from each sub-part by passing it through a Siamese-based network, namely the Siamese Multi-Scale Network *(SMSNet)*.

In the second approach, we come up with a hierarchical classification scheme to perform person re-identification in which color-based matching is initially employed to reduce the gallery set by retaining only individuals with similar appearances in the gallery set, following which Siamese Network-based matching is carried out to find the best match from this reduced gallery set. This approach has also been seen to perform re-identification from images quite accurately and efficiently, without requiring the comparison to be carried out between the input test image and each of the images present in the original gallery set through Siamese Network-based matching. We have made a thorough comparative study of the two proposed approaches on various data sets and determined the best among these. We have also compared our proposed approaches with popular existing methods and observe that both these methods perform the best among the approaches used in the comparative study. We have got a conference publication related to the first approach in the 17^{th} International Conference on Signal Processing and Multimedia Applications, and a journal publication related to the second approach in the Elsevier Journal of Visual Communication and Image Representation.

1.3.2 Effective Handling of Motion Features in Video-Based Person Re-identification

A scenario similar to that described for the previous contribution has also been used here. However, instead of performing re-identification with non-sequential image frames, here we consider the availability of person walking videos with sequential frames so that the temporal information derived from the walking pattern, i.e., gait of the person, can be fused with the spatial frame-level features to provide an improved re-identification accuracy. The scenario is diagrammatically explained in Figure 1.3 where it can be seen that sequential frames are available both for the gallery subjects and the test subject. Scenarios like this are commonly seen in public places like shopping malls, railway stations, etc., where the continuous recording of crowd movement is carried out. Since recurrent Deep Neural Networks can effectively capture the spatio-temporal information from sequential video frames, we propose using similar networks to extract effective spatio-temporal features for person re-identification. However, since, in general, video data sets are less extensive than image-based data sets, the recurrent networks may not get trained properly. To counter this, we propose fusing the spatio-temporal features

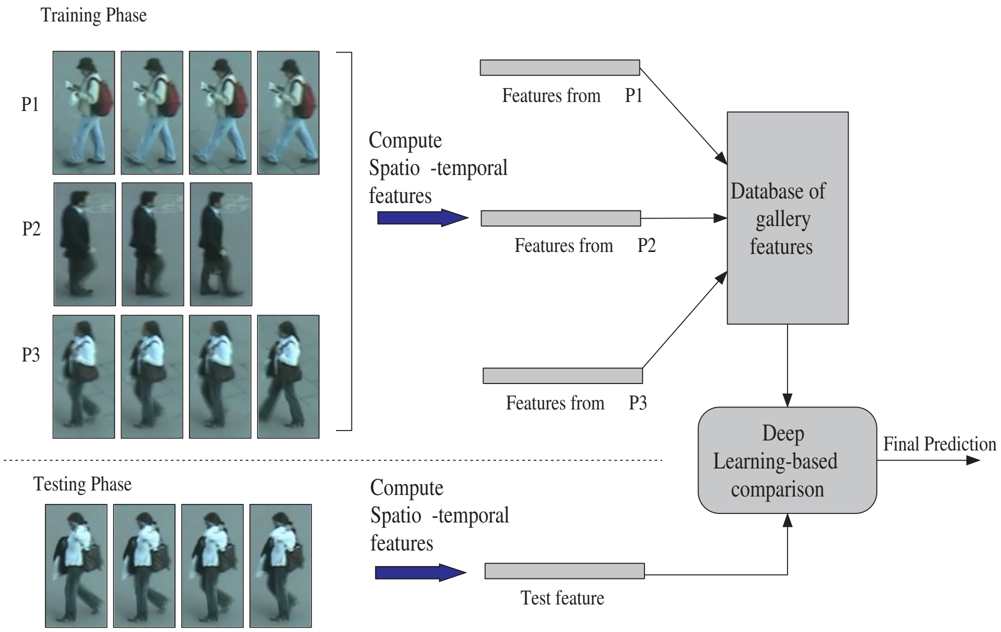
Training Phase
P1
P2
P3
Compute Spatio -temporal features
Features from P1
Features from P2
Features from P3
Database of gallery features
Deep Learning-based comparison
Final Prediction
Testing Phase
Compute Spatio -temporal features
Test feature

Figure 1.3: Scenario where spatio-temporal features extracted from sequential frames are used for re-identification

derived from the sequential frames of an input video sequence through three different recurrent networks using an ensembling approach to make a final prediction about the class of the test subject. Specifically, we use three time-series models, namely, the Full-Body Pose Attention Network *(FPAN)*, Motion Pooling Network *(MPN)*, and Long-Short Term Memory *(LSTM)* Network, each of which computes a probability of the target subject to belong to a certain class. The predictions from the above networks are finally fused to determine the correct class of the target subject. The highlights of this contribution are as follows:

- Utilizing motion information from walking videos along with the frame-level spatial features for person re-identification.
- Proposing a new ensemble model termed as Temporal Motion Aware Network (*T-MAN*) consisting of three recurrent networks to effectively extract the spatio-temporal features from sequential frames.
- Validating our approach through extensive experiments and making a compara-

tive study with existing approaches.

This work has also been published in the Springer Multimedia Tools and Applications journal.

1.3.3 Handling Occlusion in Images

Our previous two contributions are suitable for deployment only if occlusion is not present in the image frames. However, the presence of occlusion is inevitable in most real-life surveillance videos. A target subject can get partially/totally occluded due to the presence of other static and dynamic objects in the camera field of view. Our next contribution deals with proposing an effective solution to the problem of person re-identification from non-sequential image frames corrupted with occlusion. The scenario is similar to that explained with the help of Figure 1.2, the only difference being, here we consider a more unconstrained test scenario where the input non-sequential image frames may be corrupted with occlusion. As can be seen from the Figure 1.4, we have introduced an occlusion reconstruction module that predicts clean image frames

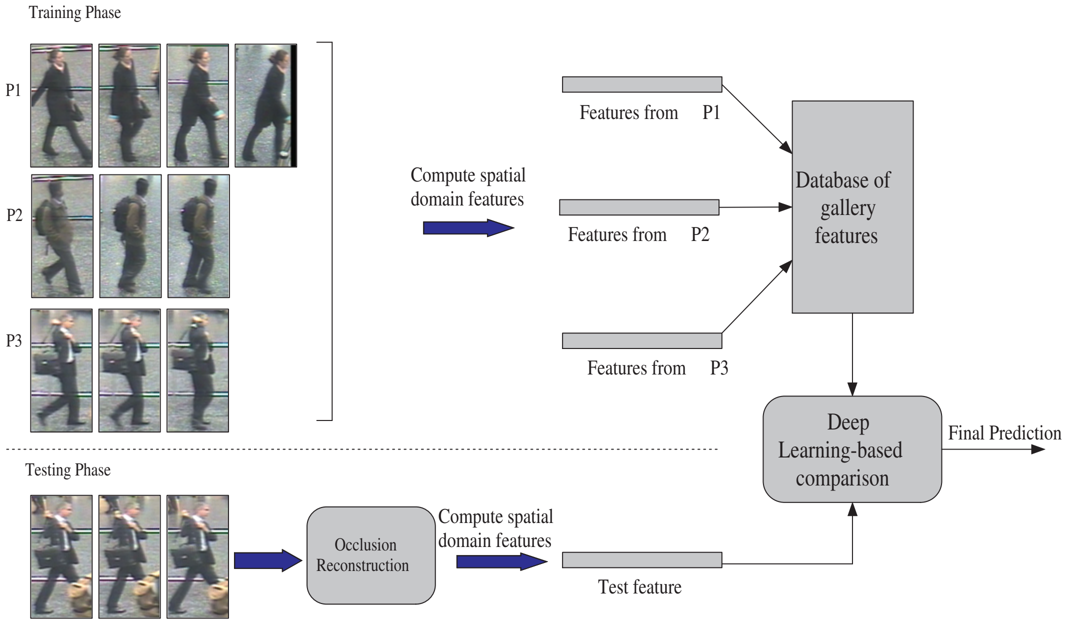

Figure 1.4: Re-identification scenario with occlusion present in non-sequential image frames

from the input occluded frames. We use Deep Neural Network-based generative models for the reconstruction phase to make the prediction robust and efficient. Once, the reconstructed frames are obtained, we carry out Deep Neural Network-based spatial-domain feature extraction and comparison, as also described for the first contribution. The important highlights of this contribution are as follows:

- Developing end-to-end Deep Learning-based pipeline to perform re-identification from occluded image frames.
- Proposing effective Deep Neural Network-based generative models to reconstruct each occluded frame from a set of non-sequential image frames.
- Training and fine-tuning the reconstruction models for effective prediction by constructing a gallery of pairs of synthetic occluded frame and its corresponding unoccluded version
- Performing extensive experiments and comparison with state-of-the-art occlusion handling methods in person re-identification

We propose effective approaches to reconstruct occlusion (if any) present in an image frame by exploiting the spatial information present in the frame before carrying out person re-identification. First, we propose a new image translation model termed as the Occlusion Handling GAN (*OHGAN*) that consists of a U-Net-type generator and a Siamese discriminator. On using this network for occlusion reconstruction, we observe that traces of occluded patches are still retained on the reconstructed images, which is mostly due to the presence of skip connections in the U-Net generator. This motivated us to study the applicability of the Convolutional *Autoencoder* (without any skip connections) in performing the occlusion reconstruction. As expected, we observe that the visual quality of reconstruction by using the Convolutional *Autoencoder* improves over the *OHGAN*. However, the images generated by Convolutional *Autoencoder* still contain noise and irregular edges. Hence, we decided to fine-tune the results further by passing the *Autoencoder*-generated images through a Deep Convolutional GAN (*DC-*

GAN), which is a widely used network for high-quality image generation. The visual quality of the reconstructed results obtained using the bi-network formed by stacking *Autoencoder* and *DCGAN* has been seen to be quite appealing, and these reconstructed images are used for re-identification. Part of this work related to occlusion reconstruction using *OHGAN* has been published in the Springer Signal, Image, and Video Processing journal. The other part of the work dealing with *Autoencoder+DCGAN*-based reconstruction has been communicated to the IEEE Transactions on Emerging Topics in Computational Intelligence and is currently under review.

1.3.4 Handling Occlusion in Videos

Next, we consider a scenario where the input frames are sequential in nature but are corrupted with occlusion as shown in Figure 1.5. Here also, we carry out reconstruction of occluded frames before the re-identification phase. However, instead of using only spatial-domain information as in the previous contribution, we exploit the spatio-temporal information from the sequential image frames as well followed by a fine-tuning

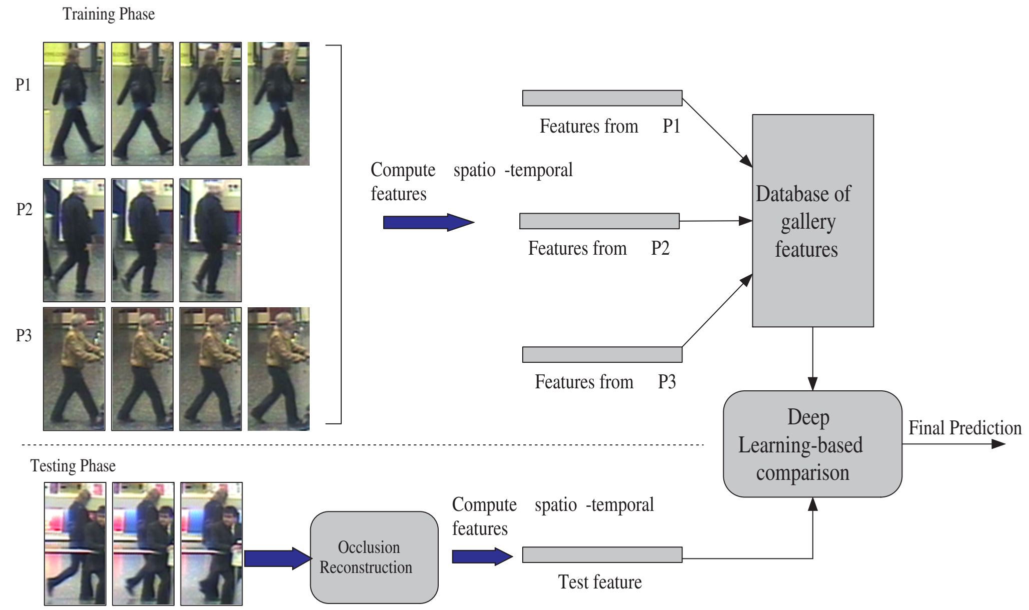

Figure 1.5: Re-identification framework for occlusion handling in sequential frames

stage to perform effective reconstruction. Finally a Deep Learning-based model is used to perform the re-identification. The important highlights of this contribution are as follows:

- Introducing a novel *Conv-LSTM*-based occlusion reconstruction by exploiting the spatio-temporal information present in the frames of a video sequence
- Preparing synthetically occluded data to serve as the ground-truth for training the occlusion reconstruction model
- Validating our approach through extensive experiments and comparative study with state-of-the-art techniques

Video-based re-identification after occlusion reconstruction by exploiting the temporal information present in consecutive frames of a video sequence has not been studied in the past. In this work, we propose an effective approach to occlusion reconstruction from videos using a Convolutional-LSTM *(Conv-LSTM)* model. As discussed for the previous contribution, here also we fine-tune the generated frames using a *DCGAN*. Since, an image frame in a video sequence is temporally dependent on the previous frames, information from the previous frames can be potentially used to predict the current frame. In a given video sequence, multiple frames (consecutive/non-consecutive) can be occluded, and reconstruction is carried out for one frame at a time by inputting the two previous frames along with the present occluded frame to the *Conv-LSTM* model. These two previous frames may be either originally unoccluded or may be predicted by the *Conv-LSTM* at a previous step. Once reconstruction of the occluded frames in the input sequence is completed, any standard classifier can be used for re-identification. Experimental results also show that the reconstruction quality of *Conv-LSTM+DCGAN* is indeed good. The work has been communicated to the IEEE Transactions on Emerging Topics in Computational Intelligence and is currently under review.

1.3.5 Constructing A New Data Set and Making the Pre-Trained Models Publicly Available

We have constructed a new data set, namely, the *IIT (BHU) Re-identification data set* to evaluate our approaches and also perform a comparative evaluation with other existing approaches. This data set consists of the front view of walking from 41 subjects in an indoor environment. It also contains a challenging situation where multiple subjects walk through the monitoring zone by wearing similar-colored clothes, thereby exhibiting similar color-based appearances. The data set is made available here, and further details on the data set have been given in Section 2.4.2 of Chapter 2. Additionally, we have made the pre-trained models out of our proposed approaches publicly available to the research community for further comparative studies here.

1.4 Organization of the Thesis

The organization of the rest of the thesis is as follows:

Chapter 2 presents a literature survey related to the traditional and recent approaches along with the description of the popular data sets used for research on person re-identification. In this chapter, we also discuss the tools and frameworks used to implement the different person re-identification algorithms presented in the thesis.

Chapter 3 describes two effective approaches for person re-identification suitable for application in unoccluded scenarios, one of which is based on multi-scale feature extraction from input images, while the other follows a hierarchical scheme that employs a different feature matching stage at each level of the hierarchy and removes vastly dissimilar elements at each level to make the search more focused. We have also made a rigorous comparative study of these two proposed methods with some popular person re-identification approaches developed for unoccluded scenarios.

Chapter 4 focuses on video-based person re-identification and describes an ensemble-based feature fusion technique to perform re-identification from sequential image frames. The Deep networks used in the ensemble are three time-series Neural Network models based on RNN. Comparative performance analysis with existing video-based person re-identification methods has also been presented here.

Chapter 5 addresses the problem of re-identification in the presence of partial occlusion in image frames. Here, we discuss robust Neural Network-based methods for both occlusion reconstruction and re-identification along with a comparative study with existing occlusion handling methods.

Chapter 6 also deals with occlusion handling in re-identification, but it considers situations where sequential image frames are available. Here, we discuss the application of a *Conv-LSTM* to reconstruct occluded frames by exploiting the spatio-temporal information from preceding video frames.

Chapter 7 concludes the thesis and points out future scopes of research in the area of person re-identification.