Chapter 2

Related work

This chapter presents a thorough literature survey on person re-identification. In Section 2.1, we provide an overview of the traditional approaches that do not use any learning technique. Next, we discuss the metric learning-based approaches and provide an elaborate discussion on the recently developed Deep Learning-based person re-identification methods in Section 2.2. Based on the literature review, we point out the limitations of the existing methods and possible future research scopes in Section 2.3. Next, we detail the publicly available data sets for person re-identification in Section 2.4 and also provide a brief discussion on the Deep Learning frameworks used in the implementation in Section 2.5. Finally, we summarize the chapter in Section 2.6.

2.1 Traditional Person Re-Identification Methods

Person re-identification has been among the most intriguing subjects in research on digital surveillance over the last decade. Erstwhile, researchers made use of inter-camera relationships to study the matching process. The traditional approaches to person re-identification broadly fall under two categories - contextual and non-contextual, which are highlighted in the following two sub-sections.

2.1.1 Contextual Methods

These methods use external contextual information to extract features for establishing correspondences between images. These usually include camera geometry information and camera calibration as context.

Using Camera Geometry: The methods that utilize camera geometry focus on using spatial and temporal information through inter-camera relations such as camera pose. This inter-camera relationship has been modeled as a probability distribution function of space-time parameters in [13]. Camera networks calibrated with entry-exit points and transition times are used in [14] to obtain a bipartite graph describing the topology of the camera network which is coupled with temporal information to realize a tempo-topographical model of the network. Unobserved regions between cameras evaluated by propagating person trajectories are studied in [15] to identify potential paths. The topology of the camera network is evaluated in [16, 17] by correlating activities across cameras with non-overlapping views. Affinity matrices modeled from spatial and temporal camera information are used in [18] to determine camera topology thus aiding the re-identification task.

Using Camera Calibration: These methods use camera calibration to extract discriminative visual descriptors for re-identification. For example, the person height is estimated by employing a homography-based 3D position estimation technique [19], while separate color and edge-energy descriptors are used for each region of a human silhouette, and finally, region-wise matching is carried out for determining the extent of similarity between the two images. Color of clothing and body build are used as additional features to extract the descriptors in [20]. The principal axis of a person is used in [21] to establish matches between individuals from different camera views and evaluate the re-identification score. In [22], a 3D point process model is presented that comprises of vertices representing appearance features like HSV histogram, mean color, the direction of normal to vertex, optical reliability of vertex, and vertex uniqueness to

measure the re-identification score. Each of the above methods requires high-resolution images for proper functioning and also detailed knowledge of the camera positions, and is hence not very suitable for application in real-life surveillance sites.

2.1.2 Non-Contextual Methods

These methods rely entirely on visual descriptors and do not incorporate any external contextual information for establishing correspondences between images. These can further be classified as Passive or Active methods, as discussed next.

Passive Methods: This category of approaches exclusively deals with visual descriptor design and does not depend on learning techniques to measure the similarities in the person appearances. The human blob extracted using color and shape features from an image is split into polar bins in [23] and a descriptor is formed from the color Gaussian model and edge pixel counts corresponding to each bin. In [24], spatio-temporal edges detected through Watershed segmentation and graph partitioning have been used to compute features for person re-identification. In [25], a two-step approach is presented that first detects the person images in a scene and next employs color features-based pictorial structures for re-identification. A weighted sum of complementary appearance features extracted from multiple camera images known as Histogram Plus Epitome (HPE) has been used for appearance matching in [26]. The work in [27] uses a weighted sum of three appearance features: weighted HSV histogram, maximally stable color regions (MSCR) [28], and recurrent highly textured local patches for re-identification. Part-based human detection using pictorial structures model followed by extraction of HSV histograms and MSCRs for each part is used for re-identification in [29], while spatial covariance detectors characterizing human body parts predicted using HOG-based detector are used to measure dissimilarity in [30] through covariance matrix distance. The authors in [31] suggest a re-identification approach in which the feature descriptors are constructed from multiple segments in a frame that are treated as multiple instances

to perform person re-identification. A histogram of Haar wavelet responses in a 4×4 region centered around Hessian interest points forms the descriptor in [32], which is used to establish correspondences among individuals by using the sum of absolute differences metric. In [33], a part-based spatio-temporal model from HS color histograms corresponding to HSV color space is developed for re-identification. Color histograms extracted from body parts detected using HOG-based detectors are incorporated into an active color model in [34]. The active color model is combined with representative meta-colors obtained through clustering to form the appearance descriptor and their weighted sum is used to compute similarity. These methods have a very fast response time, but their effectiveness is questionable in scenarios where persons wear similar colored clothes. However, in most practical re-identification scenarios, these methods can be conveniently used to get an estimate of which gallery samples closely match the appearance of the target subject.

Active Methods: This category of methods makes use of supervised or unsupervised learning techniques to extract descriptors for matching. We broadly categorize active re-identification approaches into three sub-categories: (i) color calibration-based, (ii) descriptor learning-based, and (iii) distance metric learning-based. The methods that aim to model the color relationship between cameras fall under the color calibration-based category. Among the approaches in this category, in [13], brightness transfer function (BTF) has been used to model the appearance changes between objects captured by the two cameras. BTFs have also been used in [35] by Porikli et al. to perform camera color calibration using a Dynamic Programming-based shortest path-finding algorithm. In another work [36], a cumulative computation method is presented where the BTFs are computed through the 'mean' operator.

Descriptor learning methods employ a discriminative weighing scheme for multiple features or follow a bag-of-features approach for the generation of descriptive dictionaries of features during the learning stage. For example, in [37], shape and appearance con-

text models have been considered to form the descriptors for matching. Code words describing appearance based on HOG features computed in log-RGB space are used to construct appearance descriptors using their spatial occurrence in [38], while appearance words based on SIFT and average RGB color are used in combination with group context to describe the appearance in [39]. A pair of group descriptors, one describing ratio information of appearance words within regions centered on the group, and the other containing local spatial information between labels, serve as a contextual cue for person re-identification. Adaboost learning has been used in [40] to perform re-identification by learning discriminative features from several weak classifiers. In another work [41], boosting is applied on top of covariance-based detectors derived from RGB information and Haar features to select the discriminative features which are subsequently used for person re-identification. Apart from these, Haar-like features and dominant color descriptors have also been used for person re-identification in [30]. A binary SVM classifier is employed in [42] to solve the varying camera output problem by learning the camera-pair-specific variations in the feature space formed by concatenating individual appearance descriptors. Descriptors belonging to the same person are treated as positive class samples while those belonging to different persons are treated as negative class samples.

Distance metric learning-based approaches have also been used in the literature to achieve improved re-identification performance. For example, the large margin nearest neighbor (LMNN-R) classifier [43] is one such model that learns to minimize the distance between true matches and maximize the distance between false matches. Another similar metric learning approach termed as the relative distance comparison (RDC) has been presented in [44] in which a logistic function-based comparison model is formulated for feature quantization to improve the performance of the distance metric learning. In [45], the authors proposed the extraction of biologically inspired features (BIF) from Gabor filters, and feature similarity computation using a covariance distance learning

approach. Active re-identification approaches also make use of descriptor learning that involves learning the most discriminative set of features or applying a weighting scheme to prioritize multiple features. Dimension reduction by a discriminative weighting of color, texture, and edge features, as given in [41], uses a partial least square (PLS)-based approach to come up with a reduced descriptor for person re-identification. Active methods perform more robustly than passive methods across varying re-identification scenarios, but due to the incorporation of certain learning strategies, their response time is somewhat higher than the passive approaches.

2.2 Modern Approaches

In this section, we review the re-identification methods that use modern sophisticated learning techniques. These learning-based approaches can be broadly categorized into three groups: (a) initial Deep Learning models discussed in Section 2.2.1, (b) Siamese architecture-based models discussed in Section 2.2.2, and (c) Generative Adversarial Network (GAN)-based models discussed in Section 2.2.3.

2.2.1 Deep Learning-based Approaches

To date, several Deep Learning-based person re-identification techniques have been developed by researchers worldwide. The first among these is the work in [3] that introduces a popular dataset known as *CUHK_03* and comprises of two major components: (i) a filter pairing neural network (FPNN) to handle misalignment, occlusion, and background clutter and (ii) a Convolutional Neural Network (CNN) for deep feature extraction. A deep framework for scalable distance-driven learning is proposed in [46] which uses relative distance comparison for person re-identification. In [47], McLaughlin et al. first introduced the concept of modeling temporal information in frame sequences, i.e., clips, by employing a Recurrent Neural Network (RNN). Here, the average of each RNN cell output has been used as a clip-level representation. Like [47],

Yan et al. [48] also employed RNN to encode sequence features and considered the last hidden state to preserve the entire video information. Liu et al. [49] presented a Quality Aware Network (QAN), which is an attention-weighted average to compute temporal features where the attention scores are created from the frame-level feature maps. The approaches described in [50] and [51] extract attention features as well as temporal RNN-based features to preserve the dynamic motion information. The two-stream network presented by Chung et al. in [52] computes features from both RGB images and optical flow and uses simple temporal pooling to aggregate the feature descriptors. In [53], a classification ensembling approach is discussed by fusing several Deep Networks to improve the generalization. Another end-to-end trainable architecture, namely, the Accumulative Motion Context (AMOC) has been proposed in [54] to jointly handle the appearance representation and motion context present in a given video sequence. In [55], an unsupervised approach for label estimation is presented based on a Dynamic Graph Matching (DGM) framework to improve the label estimation process in person re-identification. Here, intermediate labels have been used to iteratively refine the graph structure for labeling the data. The works proposed in [56] introduced a Harmonious Attention CNN (HA-CNN) model to jointly learn soft and hard region pixels in an input frame. In another work [57] a mask guided technique is proposed with triplet loss function.

In [58], it is assumed that all images of a person within a particular camera range lie in the same low-rank sub-space, and based on this assumption, a non-negative low-rank and sparse graph is learned to represent silhouette appearances. Next, *NCut* is employed for segmenting the silhouette to perform sub-space clustering and obtain representative features of a subject corresponding to each view. Finally, a cross-view quadratic discriminant analysis scheme is used to find the correspondences between the subjects in the two cameras. Another multishot-based person re-identification technique is described in [59] in which reference points-based similarity metric is used for

pedestrian re-identification. An unsupervised approach based on transfer learning of spatio-temporal features is described in [60] in which a classifier is trained to learn the spatio-temporal features from the input frames. These learned features are next compared with the spatio-temporal features obtained from the test data using a Bayesian fusion model to perform re-identification. The multi-scale feature extraction approach described in [61] deals with differently-scaled images of the same person and determines the most suitable scale for matching. The approach in [62] aims to maximize the total number of correct matches in a camera network. Here, the authors have used a pre-trained CNN model for feature representation, and next compute a similarity matrix between each pair of subjects using a Cosine similarity metric. A Gradient Descent algorithm is followed to maximize the global similarity while simultaneously minimizing the constraints computed based on the similarity matrix, inter-camera inconsistencies, etc. In another recent work [63], a graph-based matching framework is introduced for video-based person re-identification, termed as the Multi-Granular Hypergraph in which hypergraphs are constructed from multiple spatial part-based features across an input video-sequence to perform re-identification.

2.2.2 Siamese Network-based Approaches

Siamese network was introduced in the early '90s to identify the similarity/dissimilarity between two input images. Unlike, other Neural Network-based classification models, in the case of the Siamese network, there are two parallel channels of identical feature extractors with the same parameters and weights followed by a feature similarity layer that provides a measure of similarity between the two images. To perform re-identification of a test subject using a Siamese model, the similarity scores have to be computed between all possible pairs of the test image and gallery image, and the class of the gallery image with the maximum similarity is assigned as the class of the test subject. The first Siamese neural network was proposed in [64] and later on, with the

development of Deep Learning architectures, there have been significant advancements in this area. Siamese networks have also been extensively used in research on person re-identification in the past [1, 52, 65–69]. Among these, the deep metric learning introduced by [65] for person re-identification uses a Siamese network to compare the color and texture features. In this work, the authors also perform a cross-database experiment to test the robustness of the approach in practical scenarios. An improvement to the network architecture used in [65] has been proposed in [1], in which four convolutional layers are used to obtain deeper features. The work in [66] focuses on developing a partition-based appearance model for learning. Here, a Siamese Convolutional Neural Network (SCNN) is employed to compute features from different body parts, and finally, Cosine similarity is used to compute the matching score. The architecture proposed in this work is capable of learning both the rigid and latent body part appearances. Additionally, different dilation rates impart a higher degree of robustness to the network but reduce its efficiency significantly. The work in [67] describes a Siamese network with convolutional layers to establish multi-level similarity perception by incorporating different similarity constraints to both low-level and high-level feature maps. In another work [52], a Siamese network is employed for video-based person re-identification in which a weighted two-stream objective function has been used to combine the spatial and temporal features from the two branches of the Siamese network to predict correspondences between individuals in the two cameras. However, the use of a single model to learn both the complex spatial and temporal features may not be effective especially if the re-identification scenario is complicated with varying subject poses and background. Rather, feature fusion from multiple models through an ensembling approach is likely to capture the different aspects of the spatio-temporal motion information contained in the video frames, thereby improving the re-identification accuracy. In [68], compact and style invariant representation issues are solved by introducing multiple classifiers in a Siamese network. In another similar approach reported in [69],

the compact and style invariant representation issues in person re-identification are addressed by introducing a new attention-driven Siamese learning architecture, namely, Consistent Attentive Siamese Network.

Siamese networks have been widely used for imbalanced data classification, and these are also capable of performing few-shot learning with high accuracy [70, 71]. Training a Siamese network properly requires only an extensive data set of positive and negative pairs of training samples, and there is no requirement of the presence of a large number of samples from each subject. Since in case of person re-identification, we determine identity correspondences from images/videos captured by two cameras, usually, the number of training data per sample is quite less, which is insufficient for training a Deep non-Siamese classifier effectively. Due to this reason, Siamese architectures have gained immense popularity in research on person re-identification and have also been used in our proposed approaches discussed in the subsequent chapters.

2.2.3 GAN-based Approaches

Generative Adversarial Networks (GANs) are highly effective in performing various types of image translation and prediction tasks [72–74], and also play an important role in reconstructing the missing information present in the occluded image frames. These networks have also been used in the past to perform re-identification after GAN-based reconstruction of the frame embeddings. For example, in [75], spatio-temporal features are used to come up with a model that can jointly perform occlusion reconstruction in the encoded space and re-identification. The accuracy of this method is not appreciably high and this is likely to improve if two separate sub-networks are used dedicated towards occlusion reconstruction and re-identification. Qian et al. in [74] developed a Pose-Normalization GAN (PN-GAN) to generate composite images of an individual with a given pose translated to another target pose. However, modern Deep Learning models are capable of performing accurately even without employing the pose transfer

phase. In [73], another solution to handle pose variation has been discussed that makes use of the Feature Distilling GAN (FD-GAN). This is a Siamese architecture with multiple discriminators to learn identity-related and pose-unrelated representations. In [72], cross-modality has been tested with person re-identification where a Cross-GAN architecture has been used to make predictions using features extracted from RGB and infrared images. In [76], a GAN-architecture, namely, the Person Transfer Generative Adversarial Network (PTGAN) is discussed that focuses on performing re-identification by handling the lighting variations, viewpoint, and pose changes in the person re-identification task. In another work [77], Multi-Camera Transfer GAN is proposed to improve the re-identification performance for cross-data set experiments. In [78], Zhang et al. presented the Part-based Non-Direct Coupling Embedded GAN to perform re-identification by incorporating a block-based learning technique. In a different work [79], the problem of person re-identification from low-resolution images has been addressed in which a GAN model is employed to translate the input low-resolution images into equivalent high-resolution images before performing person re-identification.

Next, we highlight some important recent work on person re-identification by providing a brief summary of each method along with its limitations.

Table 2.1: Summary of some recent approaches to person re-identification

Ref. No.	**Conference/ Journal Name (Year)**	**Summary**	**Limitations**
[56]	IEEE Conference on Computer Vision and Pattern Recognition (2018)	Introduced Harmonious Attention CNN (HA-CNN) model for joint learning of soft pixel attention and hard regional attention	Not suitable for Occluded image frames

[80]	IEEE/CVF International Conference on Computer Vision (2019)	Proposed an Attentive but Diverse Network (ABD-Net). that seamlessly integrates attention modules and diversity regularizations throughout the entire network to learn features	Not robust for large Data sets
[81]	IEEE Transactions on Circuits and Systems for Video Technology (2020)	Proposed the feature refinement and filter network to remove background interference	Not suitable for Video sequences
[82]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)	Proposed multi-label classification for unsupervised person Re-ID	Inefficient in exploiting spatio-temporal information
[83]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)	Presented an augmented discriminative clustering (AD-Cluster) technique that estimates and augments person clusters in target domains	Efficient but not so robust in case of large data set
[84]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)	Addressed the missed contextual cues by exploiting both the accurate human parts and the coarse non-human parts	Temporal features are not considered
[57]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)	Introduced a novel region-level triplet loss to restrain the features learnt from different regions	Doesn't work well for Occluded image frames

[85]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)	Proposed a Group-aware Label Transfer (GLT) algorithm, which enables the online interaction and mutual promotion of pseudo-label prediction and representation learning.	Doesn't work well for pose variation and low resolution images
[86]	IEEE/CVF International Conference on Computer Vision (2021)	Proposed an end to end Part Aware Transformer (PAT) for occluded person re-identification	Not suitable for sequential image frames
[87]	IEEE/CVF International Conference on Computer Vision (2021)	Presented a large scale unlabeled person re-identification dataset "LUPerson" and make the first attempt of performing unsupervised pre-training	Not generic for other publicly available data sets
[88]	AAAI Conference on Artificial Intelligence (2020)	Proposed multi-label classification for unsupervised person Re-ID	Inefficient in exploiting spatio-temporal information
[89]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)	Proposed an effective Relation-Aware Global Attention (RGA) module which captures global structural information for better learning	Efficient but not so robust in case of large data set

[90]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)	Overcome the difficulty of lack of suitable dataset, by collecting a small yet representative real dataset for testing whilst building a large realistic synthetic dataset for training	Sequential dataset is not considered for feature extraction
[91]	28th ACM International Conference on Multimedia (2020)	Proposed a coupling optimization method including the Domain-Invariant Mapping (DIM) method and the Global-Local distance Optimization (GLO)	Doesn't work well for Occluded image frames
[92]	Association for the Advancement of Artificial Intelligence (AAAI) (2021)	Proposed to split each single cluster into multiple proxies and each proxy represents the instances coming from the same camera	Not an end to end framework for person re-identification
[93]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)	Proposed a Global-guided Reciprocal Learning (GRL) framework for video-based person Re-ID	Not robust for large Data sets
[94]	IEEE Transactions on Neural Networks and Learning Systems (2021)	Proposed to incorporate the pose information into the re-id framework, which benefits the model	Not scalable for large scale data sets sequences

[95]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)	Proposed a simple yet effective method that is robust to sparse and noisy pose information	Inefficient in exploiting spatio-temporal info-rmation
[96]	IEEE Transactions on Image Processing (2021)	Proposed a joint learning framework to learn better feature embeddings via high precision neighbor pseudo labels and high recall group pseudo labels.	Doesn't work for spatio-temporal information
[97]	IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)	Proposed an occlusion-robust block, Region Feature Completion (RFC), for occluded reID that can recover the semantics of occluded regions in feature space	Temporal features are not considered
[98]	Association for the Advancement of Artificial Intelligence (AAAI) (2021)	Paper presents Matching on Sets (MoS), a novel method that positions occluded person re-ID as a set matching task without requiring spatial alignment	Not suitable for reconstructing sequential frames
[99]	IEEE Transactions on Image Processing (2020)	Proposed part segmentation as an assistant body perception task during the training of a ReID model	Not suitable for Occluded image frames

[100]	IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)	Proposed a novel Spatial-Temporal Graph Convolutional Network (STGCN) to solve temporal and spatial modelling jointly	-
[101]	IEEE Transactions on Image Processing (2020)	Proposed an auto-encoder model that can be plugged into any deep network to mine latent information in an unsupervised manner	Not suitable for exploiting spatio-temporal information
[102]	IEEE/CVF International Conference on Computer Vision (2019)	Mainly focus on how to learn view-invariant features by getting rid of view specific information through a view confusion learning mechanism.	Not robust for large data sets
[103]	AAAI Conference on Artificial Intelligence (2020)	Proposed an approach called Viewpoint-Aware Loss with Angular Regularization (VA-reID) to handle underlying relationship between different viewpoints	Efficient but not so robust in case of large data set
[104]	AAAI Conference on Artificial Intelligence (2020)	This study argue that by explicitly adding a sample filtering procedure after the clustering, the mined examples can be much more efficiently used	Not valid for varying illumination conditions

[105]	AAAI Conference on Artificial Intelligence (2020)	Introduced Semantics Aligning Network (SAN) which consists of a base network as encoder (SA-Enc) for ReID, and a decoder (SA-Dec) for reconstructing/ regressing the densely semantics aligned full texture image	Semantic alignment works fine but somewhat compromising for re-id accuracy

2.3 Scopes for Further Research

Based on the extensive literature review, we point out the scopes for work in the area of Computer Vision-based person re-identification as mentioned point-wise next:

- Existing methods to person re-identification using Deep Learning perform accurately but are computationally time-intensive. Hence, these are not effective enough for handling large gallery sets. On the other hand, traditional passive approaches employing appearance-based features are fast but cannot perform robustly in all situations such as in presence of varying illumination or pose changes. Since passive approaches suit well in shortlisting the closely-matched samples from the gallery set, an effective combination of a suitable passive method with a modern Deep Learning-based technique may help in achieving an improved accuracy at the cost of less processing time, which has not been given significant attention in the past.
- Developing time-efficient video-based re-identification from spatial and temporal information has also not received much focus. A few video-based re-identification methods that exist in the literature are time-intensive due to employing expensive operations such as optical flow followed by temporal pooling to derive the motion features. Also, existing data sets to video-based re-identification usually contain

non-sequential images of persons in varying poses from different viewpoints, which does not typically resemble the manner in which a surveillance camera captures images/videos. This is since the images present in the existing data sets do not correspond to any sequential activity such as walking, running, etc. The recent video-based re-identification approaches are also not focused towards extracting effective kinematic features corresponding to some temporal activity. Rather, these focus more on computing appearance-based features through some spatial descriptors corresponding to the frames in the sequence.

- Research on person re-identification from images/videos in the presence of occlusion is still not mature enough and the existing methods are not effective enough in reconstructing the occluded frames from image/video data. There are significant scopes for improving the existing literature by developing effective methods for occlusion reconstruction by exploiting the spatial information from images and spatio-temporal information from videos.
- The open-set re-identification scenario is a challenging task in which the gallery set evolves continuously. However, majority of the approaches developed to date focus on closed set re-identification. In future, attention needs to be given to develop effective methods for open set re-identification as well.

In this thesis, we extend the state-of-the-art by proposing novel approaches to person re-identification to address the first three scopes mentioned above. In Chapter 3, we present effective image-based person re-identification approaches, one of which carries out multi-scale feature extraction for improved re-identification, whereas the other combines a passive color-based matching technique with Siamese network-based matching in a hierarchical manner to determine the correct class. In Chapter 4, we propose an ensemble of three different time-series models, namely, RNN, LSTM, and GRU to perform person re-identification effectively using the spatio-temporal information present in the video sequences. Chapters 5 and 6 discuss our proposed approaches towards

handling the challenging occlusion scenario in person re-identification from images and videos, respectively. While in Chapter 5, we describe a plausible solution to occlusion reconstruction using spatial information from image data, in Chapter 6, we extend the previous approach to reconstruct occluded video frames using the spatio-temporal information from the preceding consecutive frames. All our proposed approaches have been validated through extensive experiments as well as through comparative study with popular existing techniques. The trained models have also been made publicly available to the research community for further comparison here.

Next, we provide detailed descriptions of the data sets for person re-identification used in the thesis along with the tools and frameworks used to develop the Deep Learning models. The evaluation metrics used to test the proposed re-identification and reconstruction algorithms to be discussed in the subsequent chapters, i.e., Chapters 3 to 6 have also been stated in this chapter.

2.4 Data Set Description and Evaluation Metrics

In this section, we explain the details of the data sets used in the study. As already discussed, there can be two possible categories of data sets, namely, (i) image-based data sets and (ii) video-based data sets. The data sets used in the thesis to evaluate our proposed algorithms are described by grouping these into the appropriate categories in the following two sub-sections.

2.4.1 Image-based Re-Identification Data Set

These data sets consists of either a single image frame or multiple non-sequential image frames corresponding to each subject. We have used five public data sets belonging to this category, namely, *VIPeR* [106], *CUHK_01* [107], *CUHK_03* [3], *Market1501* [108], *Occluded ReID* [11], *Partial ReID* [109], and *Partial iLIDS* [110].

VIPeR **[106]** stands for Viewpoint Invariant Pedestrian Recognition data set. As

the name suggests, the data set contains images of individuals captured from different viewpoints. There are 1264 images from 632 persons with each person having exactly two images from two different viewpoints. The images in this data set are scaled down to 128×48 pixels.

CUHK_01 **[107]** data set is captured at the Chinese University of Hong Kong. There is a total of 3884 images from 971 different individuals. The data set consists of human silhouette images that are cropped and normalized to fixed dimensions of 168 pixels height and 60 pixels width.

CUHK_03 **[3]** data set has been also captured at the Chinese University of Hong Kong and is more extensive than the *CUHK_01* in terms of the number of subjects and cameras used during data capture. This data set consists of 13160 images from 1360 individuals captured from six different viewpoints. Each individual is observed from two disjoint camera views. On average, there are five to eight images per person corresponding to each view. The data set contains three cells in a *.mat* file, tagged as 'detected', 'labeled', and 'testsets'. This data set provides two types of annotations, the 'detected' cell contains the bounding box information of each silhouette predicted using a pedestrian detector, whereas in the 'labeled' cell the bounding boxes are labeled by humans. The 'testsets' cell provides 20 different test sets each of which contains 100 different test-ids.

Market1501 **[108]** data set is collected in an open environment at Tsinghua University. This data is collected with six overlapping camera views, and among these cameras, five capture high-resolution images, and one capture low-resolution images. In total, there are 32268 images from 1501 individuals, out of which 12936 images are marked as training images and 19732 images are marked as test images. This data set is quite extensive as well as challenging due to its large size and variability.

Occluded ReID [11] data set is an image-based data set captured using mobile cameras. The data set contains only non-sequential images which are arranged in two

folders, namely, gallery and query images. Each of these folders has 1000 images from 200 different individuals. While the gallery folder contains non-occluded images only, all the images present in the query folder are occluded.

Partial ReID [109] data set is a partial re-id benchmark and holds 600 images from 60 persons. Like the Occluded ReID data, this data set is also arranged in two sub-folders, each containing five images per person with and without partial occlusion.

Partial iLIDS [110] is also a partially occluded re-identification data set based on the original iLIDS data set [111]. This data set is captured by multiple cameras in the airport and contains only 238 images from 119 persons. Each of the gallery and query sets contains 119 images of unoccluded and synthetically occluded images. Table 2.2 provides an overview of the important features of each of the image-based data sets explained above. It may be noted that although the data sets *Occluded ReID*, *Partial*

Table 2.2: Image-based re-identification data sets

Dataset Names	Number of Cameras	Number of Images	Number of Identities
VIPeR [106]	2	1264	632
CUHK_01 [107]	2	3884	971
CUHK_03 [3]	5 pairs	13160	1360
Market1501 [108]	6	32268	1501
Occluded ReID [11]	-	1000	200
Partial ReID [109]	-	600	60
Partial iLIDS [110]	-	238	119

ReID, and *Partial iLIDS* provide occluded images from all the identities, these do not contain unoccluded frames corresponding to the occluded frames. This makes it difficult to train any Deep Learning-based occlusion reconstruction model that requires occluded data and the corresponding unoccluded data as well. Hence, in our work, we could not use the occluded samples originally present in the data sets for training our reconstruction models. Rather, we synthetically occlude the gallery sets and use these along with the corresponding unoccluded images for training purposes. The process of generation of synthetic occlusion will be discussed in more detail in Chapter 5.

2.4.2 Video-based Re-Identification Data Set

These data sets are constructed from sequential image frames (or, videos) of each subject. We have used four publicly available data sets under this category, namely, *PRID-2011* [112], *iLIDS-VID* [111], *MARS* [113], and a new data set captured in our laboratory, namely the *IIT(BHU) Re-Identification Data Set*.

***PRID-2011* [112]:** This data set consists of images from 749 persons captured by two cameras with non-overlapping fields of view, and 200 individuals among these appear in both the camera views. The images contained in this data are captured in non-crowded regions with rare occlusion and in presence of a relatively clear background. For our experiments, we consider only the common set of *200* subjects that appear in the fields of view of both the cameras.

***iLIDS-VID* [111]:** This data set consists of pedestrian images captured in an airport arrival hall. It is constructed from two non-overlapping camera views and contains 600 image sequences from 300 distinct individuals. This data set incorporates more challenging scenarios compared to that of *PRID-2011* by considering occlusion, background clutter, viewpoint, and lighting variations, etc. The number of frames present in the video sequences in this data set ranges from 23 to 192 with an average of 73.

***MARS* [113]:** This data set is the largest video re-identification data set to date. It consists of about *20000* video sequences from *1261* individuals. Each of the sequences is obtained automatically by using the Deformable Part Model [114] detector. The tracking of individuals is carried out through the GMMCP [115] tracker. In this data set, video sequences of each person are captured by a minimum of two cameras and a maximum of six cameras. On average, it contains 13 video sequences for each person.

***IIT(BHU) Re-Identification Data Set*:** As explained in Section 1.3 of Chapter 1, one of the contributions of the thesis is constructing a video-based re-identification data set from multiple subjects in which the consecutive frames will depict some form of temporal activity such as walking or running, and not just person images with varying

poses captured from different viewpoints. Hence, we construct an indoor data set in the Pattern Recognition Laboratory of the Computer Science and Engineering Department, IIT(BHU) Varanasi using two cameras each of which captures the front view walking video of 41 different subjects. Two different scenarios are considered while preparing the data set: (i) a general scenario in which no constraint is imposed on the clothing condition of the subjects but the clothing of a particular subject remains the same in the two camera fields of view, and (ii) a more challenging scenario in which all the subjects wear similar-colored clothes during the video capturing phase. For the second scenario, we use a smaller set of 20 individuals out of the previous set of 41 subjects. The average number of frames per person in the data set is 48 and the complete size of the uploaded data set is 65.1 MB. The data set can be downloaded by clicking here. The important features of the different video-based person re-identification data sets discussed above are presented in Table 2.3. It may be noted that since a video can be

Table 2.3: Video-based re-identification data sets

Data Set Names	**Number of Cameras**	**Number of Images**	**Number of Identities**
PRID-2011 [112]	2	24541	749
iLIDS-VID [111]	2	42495	300
MARS [113]	6	1191003	1261
IIT(BHU) Re-Identification Data Set	2	1963	41

viewed as a sequence of image frames, any re-identification data set meant for developing a video-based re-identification technique can be used for implementing an image-based re-identification algorithm as well.

2.4.3 Evaluation Metrics

The performance of the approaches described in Chapters 3 to 6 have been evaluated using popular evaluation parameters, namely, Accuracy, Cumulative Matching Characteristics (CMC), Mean Average Precision (map), and Dice Similarity Coefficient

(DSC). The overall re-identification performance has been evaluated using Accuracy, *CMC*, and/or *map*, while the reconstruction models discussed in Chapters 5 and 6 have been evaluated using *DSC* metric.

Accuracy: Accuracy (also termed as Rank 1 accuracy) of a re-identification model is defined as the percentage of subjects correctly re-identified by the model.

Cumulative Matching Characteristic *(CMC)*: The Cumulative Match Characteristic (CMC) curve measures the rank-wise improvement in prediction accuracy of the re-identification model. In most real-life problems, obtaining the correct class at the topmost prediction of the classifier, i.e., at Rank 1, is not always required. Rather, it is sufficient if the correct class falls within the top few predictions of the model. Here, CMC has been used to provide a graphical representation of the improvement of re-identification accuracy with the increment in the rank value and has been extensively used in several studies to evaluate the effectiveness of a method as well as perform a comparative study among different methods.

Mean Average Precision *(map)*: The mean average precision *(map)* is another popular metric for identification tasks. In this metric, the average precision is calculated for each class and averaged based on the total number of classes to obtain the *map* score.

Dice Similarity Coefficient *(DSC)*: This metric is used to evaluate the effectiveness of the reconstruction models explained in Sections 5.2.1, 5.2.2, and 6.1 in later chapters. This metric computes the degree of spatial overlap between the predicted image and the desired ground-truth image and is defined as:

$$DSC = \frac{2 * Area\ of\ Overlap\ in\ Two\ Images}{Total\ No.\ of\ Pixels\ in\ both\ the\ Images}. \tag{2.1}$$

2.5 Tools and Frameworks

Here, we explain the tools and frameworks used to implement the re-identification approaches described in Chapters 3 to 6. There exist several popular frameworks to

implement Deep Learning models. Among these, we have primarily used open-source Pytorch, Tensorflow, and Keras.

2.5.1 Keras

Keras [116] is open-source software released in March 2015 for implementing Deep Learning models. It provides a Python interface to the Artificial Neural Networks and Tensorflow. It is designed for fast experimentation with Deep Neural Networks and is user-friendly, extensible, and modular. It is first written and maintained by Google engineer François Chollet, who is also the author of the XCeption Deep Neural Network model [117]. Apart from standard Deep Learning flexibility, it also provides a platform for users to distribute their Deep Learning models on smartphones and the web.

2.5.2 Tensorflow

Tensorflow [118] is also a free and open-source software library developed by the Google Brain team in November 2015 for Machine Learning applications. It is an extensive library with several useful functions and can be used for a variety of Machine/Deep Learning related tasks. This library focuses particularly on providing an interface to distribute Deep Learning models. It is a symbolic math library based on Data Flow and Differentiable Programming. Tensorflow is now being used all over the world by various companies and research communities (DeepDream [119]) to implement Deep Learning models.

2.5.3 Pytorch

Pytorch [120] is a free-to-use open Machine Learning library released in September 2016. It is based on Torch which is a scientific computing framework and scripting language based on Lua programming language. PyTorch is useful for Computer Vision and Natural Language Processing and is primarily developed by Facebook's AI Research

lab (FAIR) [120,121]. Pytorch provides two important features: Tensor computing (like NumPy) with strong acceleration via graphics processing units (GPU) and Deep Neural Networks built on a type-based automatic differentiation system.

2.6 Summary

In this chapter, we have reviewed the existing approaches to person re-identification starting from the initial contextual and non-contextual methods to the modern Deep Learning-based approaches. Among the non-contextual approaches, passive methods compute appearance-based descriptors and perform quite efficiently. Although the feature extraction process of the passive methods is quite simple, these are not so effective in situations where the appearances of subjects are similar. Active learning methods based on distance metric learning and descriptor learning perform more robustly than the passive methods and thus have received higher attention in the past compared to passive approaches. In recent years, more focus has been given to developing Deep Learning-based methods for person re-identification. Among these, Siamese network-based re-identification algorithms have been seen to be significantly effective and are being extensively used in several studies on re-identification. However, an inherent problem associated with any Deep Learning algorithm is its high response time due to the computations involved in its multiple layers. Since passive approaches can provide a good estimate regarding which individuals in the gallery set match closely with the target subject, an effective combination of a suitable passive approach with a modern robust Deep Learning-based technique may help in achieving improved re-identification accuracy at the cost of less processing time, and this requires further study. An effective combination of spatial and temporal features to perform person re-identification from videos is also another scope for research. Occlusion is a challenging problem in any real-life surveillance application. However, research on person re-identification in the presence of occlusion is not mature enough and there are significant scopes of research

in this area. Also, to the best of our knowledge, there does not exist any method in the literature that performs re-identification after reconstruction of corrupted frames in occluded video sequences.