Chapter 3

Person Re-Identification from Still Images

In this chapter, we focus on extracting effective appearance-based descriptors for image-based person re-identification. Specifically, we consider a scenario where a set of image frames of a person are only available for re-identification. These set of frames may have different poses and may be captured under varying lighting conditions. Also, these frames may/ may not correspond to any specific activity sequence, as already explained using Figure 1.2. Traditional approaches to person re-identification [30, 31, 40, 107, 122–124] extracts spatial-domain features from the available set of frames corresponding to each subject, and are suited for deployment in the considered re-identification scenario. A limitation of this category of approaches is that due to the use of appearance-based information these methods are usually not robust to varying lighting conditions or pose changes of a subject. With the advent of Deep Learning, research has been also done to study if application of Deep Neural Networks can improve the accuracy and robustness of the primitive approaches further [1, 12, 86, 90, 92, 98, 125–127]. However, Deep Learning architectures usually involve a large set of parameters which makes the process time-intensive, and also the color information of an input image gets faded out

as the image is gradually passed across the layers of the Deep Network. Also, similar to traditional approaches, the Deep Learning-based methods fail to perform well if the scale of the images captured during testing is drastically different from that of the images present in the gallery set. Since color is an important appearance cue for person re-identification, it appears that effective fusion of the traditional appearance-based approaches with the modern Deep Learning approaches can lead to a more accurate and robust prediction.

In this chapter, we study the effectiveness of using two different feature extraction techniques, and next make a comparative analysis to determine the best among the two. Our first approach is based on multi-scale feature extraction through dilation applied at the layers of a deep network to be discussed in Section 3.1, whereas the second approach is based on a two-step hierarchical classification scheme to be discussed in Section 3.2. In the second approach, dominant color information has been used at the first level of hierarchy to eliminate individuals with significant differences in their appearances, thereby enabling the search to focus on only those individuals that closely match with the target subject in terms of appearance features and prevent it from getting biased towards a completely different element in the search space. Next, a Siamese network is used to identify the correct match from the reduced set of samples.

The surveillance setup for re-identification considered in this chapter is explained using Figure 3.1. The figure shows a multi-camera surveillance setup in which there are designated entry and exit points. Each of the n cameras (labeled as Cam_1, Cam_2, ..., Cam_i, ..., Cam_n) shown in the figure captures walking videos of persons from the front view. This scenario is similar to that found in concert and movie halls, or some meeting place, where a group of people assembles in a hall, and after some time each of them exits the hall one by one. For effective tracking of individuals, we propose to install cameras on top of each entry/exit point so that both the cameras capture the front view of the walking of subjects. A gallery set will be formed from all the subjects captured

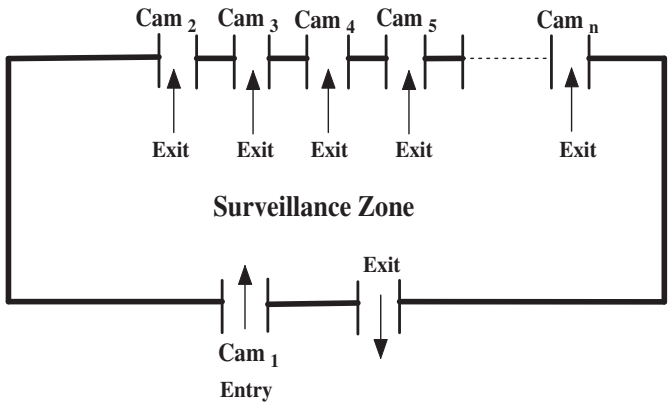

Figure 3.1: Surveillance setup for re-identification

by the several entry gate cameras, which will next be used to re-identify a subject as he/she approaches an exit gate camera. Since in the above-mentioned scenario, both the cameras capture the front view of walking, the physical appearance and clothing conditions of a subject will appear to be almost the same in the images/videos captured by the two cameras. Minor color variations may be observed due to different lighting conditions in the fields of view of the two cameras. The feature descriptors derived in this chapter are extracted using only spatial information present in the image frames.

3.1 Multi-Scale Feature Extraction for Person Re-Identification

In this section, we discuss our proposed work on multi-scale feature extraction for person re-identification. Due to the effectiveness of Siamese networks in person re-identification task, as discussed in Section 2.2.2, here we also propose to employ Siamese network-based feature extractor to compute the descriptor for each individual. Specifically, we build a new Siamese network-based model and term it as Multi-scale Siamese (*SMSNet*) architecture. The architecture detail of *SMSNet* is given in Section 3.1.1, and training of *SMSNet* is discussed in Section 3.1.2. Finally, experimental evaluation using the *SMSNet* architecture is presented in Section 3.1.3.

3.1.1 Multi-Scale Siamese *(SMSNet)* Architecture

An insight view of the proposed *Siamese Multi-scale Network* (*SMSNet*) model is given in Figure 3.2. Table 3.1 presents the detailed network configuration used in the study. With reference to Figure 3.2 and Table 3.1, the first layer of the network consists of two parallel tied convolution layers (Conv2d_0) that accepts two input images of size 60×160, and this is followed by four more convolution layers, each equipped with dilation rates of 1, 2, and 3 (Conv2d_1, Conv2d_2, Conv2d_3, and Conv2d_4). As already explained before, the application of dilation in the convolution layers helps in obtaining a multi-scale feature representation that encodes the visual characteristics of an input image by capturing low-level features. The size of the filters at every convolution layer is 3×3, except for the first layer in which the size is 5×5, and the number of filters used in each layer is the same (i.e, 32). The feature difference layer shown after all the convolution layers is used to compute the cross-input neighborhood difference [1] between the outputs of the aggregated features extracted from both the branches of the *SMSNet*.

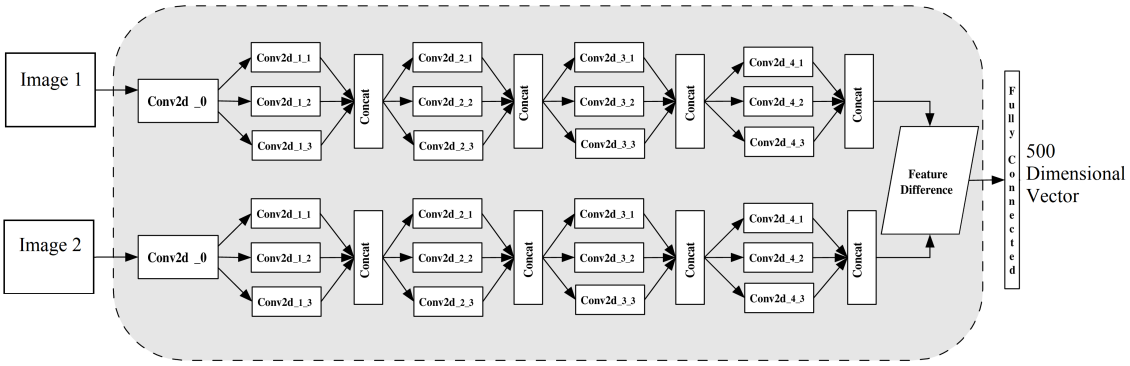

Figure 3.2: Insight view of the proposed Siamese Multi-scale Network (*SMSNet*) architecture. The first layer of convolution is unaffected by dilation parameters. All other layers are dilated with rates 1, 2, and 3, and feature aggregation has been done after each convolution layer in form of concatenation. The feature difference is computed after the fourth convolution layer.

If f_i and g_i represent the i^{th} concatenated feature maps at the final layer corresponding

Table 3.1: Layer specification of each Siamese Multi-scale Network (*SMSNet*)

Layer	**kernel**	**No. of filters**
Conv2d_0	5×5	32
Conv2d_1	3×3	32
Conv2d_2	3×3	32
Conv2d_3	3×3	32
Conv2d_4	3×3	32
Layer	**No. of neurons**	
Fully Connected	500	

to the two images input to the *SMSNet*, then the cross-input neighborhood distance $\mathbb{K}$ between f_i and g_i at each pixel location (x,y) is computed as follows:

$$\mathbb{K}(x,y) = f_i(x,y) * \mathbb{I}(n,n) - \mathcal{N}[g_i(x,y)], \tag{3.1}$$

where n is the neighborhood size, $f_i(x,y)$ is the pixel value of feature map f_i at location (x,y), $\mathbb{I}(n,n)$ denotes an identity matrix of dimensions $n\times n$ matrix, and $\mathcal{N}[g_i(x,y)]$ denotes a $n\times n$ neighborhood around pixel (x,y) of feature map g_i. In the present work, the value of n has been chosen as *5*. The use of the cross-input neighborhood distance is advantageous in the sense that it helps in obtaining the positional differences between the two input images.

3.1.2 *SMSNet* Training

As explained before, in the given problem scenario, both training and test data consist of walking sequences from the front view. Standard pre-processing techniques can be applied to segment out the bounding box containing the silhouette of the target subject and normalize the dimensions of the bounding box to a fixed height and width. Each of the feature extraction mechanisms discussed in this chapter and the subsequent chapters use these normalized cropped frames for further analysis. To capture better spatial information and preserve relative ordering among the upper, middle, and lower body parts, instead of using a single *SMSNet* to extract features from the entire frame image, we propose dividing each image into three equal segments, and pass each of these seg-

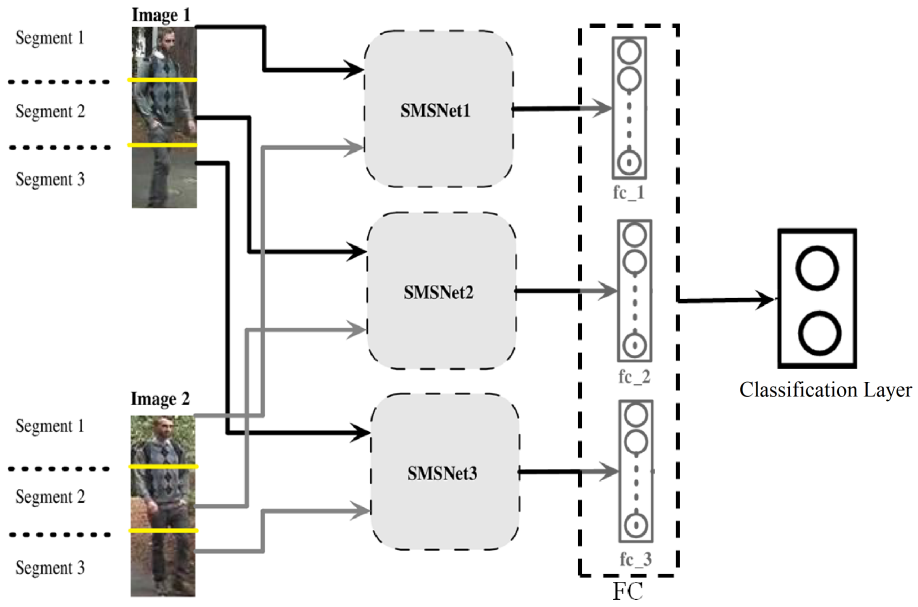

Figure 3.3: Overall framework of the re-identification approach

ments through different *SMSNet* channels as shown in Figure 3.3. These three segments are termed *Segment1*, *Segment2*, and *Segment3* in the figure, and pairs of corresponding segments from the two input images are fed to the individual *SMSNets* termed as *SMSNet1*, *SMSNet2*, and *SMSNet3*. Here, *SMSNet1* computes the cross-neighborhood distance between the first segments of the two images at its final layer denoted by fc_1, while *SMSNet2* and *SMSNet3* compute the cross-neighborhood distances between the second segments of the images and third segments of the images at their final layers denoted by fc_2 and fc_3, respectively. Each of the features in the fc_1, fc_2, and fc_3 layers is 500 dimensional, and provides useful information regarding the dissimilarity between the corresponding segments in the two input images. These features are next concatenated into a single feature vector of dimension 1500, denoted by FC. The FC layer is now fully connected with a final classification layer with two nodes representing *Similar Class* and *Dissimilar Class*, respectively. Training of the complete network is done using Adam optimizer [128] in multiple epochs by computing the binary cross-

entropy loss at the nodes of the final layer until convergence. Training of the network is done by preparing a gallery set in the form of positive and negative pairs of images. Positive pairs are formed from the images of the same identity, whereas negative pairs are formed from the images of two different identities. Each data set is divided into training, test, and validation sets. The training and fine-tuning of the *SMSNet* is done using the training and the validation sets, and the testing is done on the test set. The split information for training the *SMSNet* model will be discussed in the next section while discussing about the experimental evaluation.

3.1.3 Experimental Evaluation

For evaluating our work, we consider four data sets, namely, *VIPeR* [106], *CUHK_01* [107], *CUHK_03* [3] and *Market1501* [108]. A detailed description of each of these data sets has been given in Section 2.4.1 of Chapter 2. It may be noted that the above-mentioned data sets already provide the silhouette images extracted from the video frames. However, during working with video data in real-life scenarios, accurate localization (i.e., estimating the bounding box) of individuals in each video frame followed by frame dimension normalization has to be carried out. Since, the re-identification scenario considered in this work assumes one person to be present in the camera field-of-view at a time, localizing the moving person in the background can be done effectively using recent techniques such as [129]. Even if the bounding box detected around the moving person is not very precise, it would still not affect the re-identification accuracy much, since the proposed algorithm considers the RGB information of the entire bounding box, and does not require segmentation of clean object silhouette from the background. Hence, as long as a significant portion of the target subject appears in the estimated bounding box, our approach should be able to work satisfactorily.

We split each of the above-mentioned re-identification data set into a test set of 100 samples, a validation set of 100 samples, and a training set with the rest of the remaining

samples. The test set is used to evaluate the model during the testing phase, while the training and validation sets are used to train and fine-tune the models used for re-identification. The complete split information of the different data sets is given in Table 3.2. It may be noted that the *VIPeR* data set has a very less number of images per person (i.e., two images per person) as compared to each of the other data sets. Hence, for this data set, we use a cross-data set training strategy by fine-tuning the weights of a pre-trained Siamese model using the *CUHK_03* data set.

Table 3.2: Data set split information

Data set	No. of training samples	No. of validation samples	No. of test samples
CUHK_01	771	100	100
CUHK_03	1160	100	100
Market1501	1301	100	100
VIPeR	432	100	100

The experimental results reported in the thesis have been obtained using Tensorflow [118] on a system having 64 GB RAM, NVIDIA TITAN Xp, and NVIDIA RTX-1080Ti GPUs with a total capacity of 34 GB memory capacity. We train the proposed Siamese Multi-scale Network *(SMSNet)* model with the $l2$ regularizer using a learning rate of 0.001. To avoid over-fitting during training the network, a weight decay factor (γ) of *5e-4* is introduced at each convolution layer. The optimal values of the hyper-parameters, i.e., learning rate (η) and weight decay (γ) are determined by carrying out three-fold cross-validation using the training and validation sets for the different combinations of these hyper-parameters, and next choosing the configuration that yields the highest cross-validation accuracy. Corresponding to each data set, namely *CUHK_01*, *CUHK_03*, and *Market1501*, we consider three different combinations of η and γ namely, $C1$ (*0.01*,*2.5e-3*), $C2$ (*0.01*,*5e-4*), $C3$ (*0.03*,*5e-4*), and for each of these combinations, we perform three-fold cross-validation and observe the effectiveness of learning the training data for five different initialization of the network weights. Figure 3.4 presents the results of this experiment using box and whiskers plot. Here, each

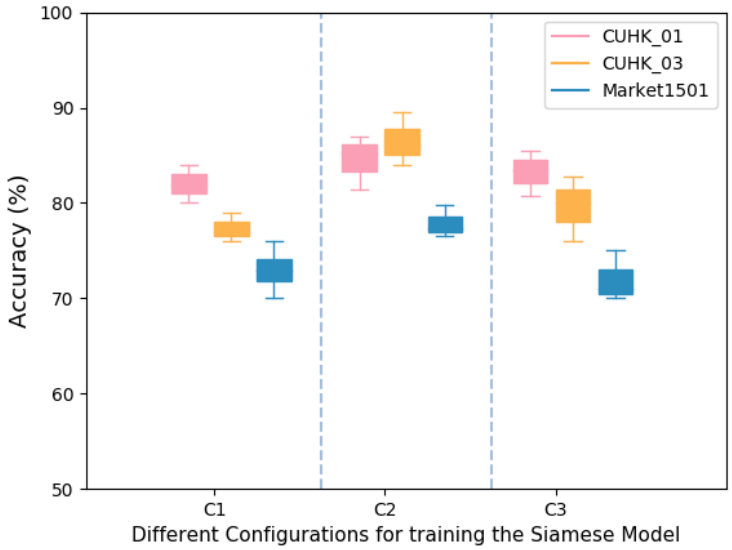

Figure 3.4: Range of three-fold cross-validation accuracy for various combinations of parameters η and γ corresponding to different the data sets by setting different initial weights of the network

box represents the variation of Rank 1 training accuracy for a particular data set (i.e., *CUHK_01*, *CUHK_03*, and *Market1501*) and network configuration (i.e., $C1$, $C2$, and $C3$). It can be seen from the figure that the inter-quartile range of boxes for the configuration $C2$ is quite less (i.e, between 1.4 to 2.5 %). The accuracy values obtained using $C2$ are also significantly higher than those obtained using either $C1$ or $C3$. Thus, the combination of hyper-parameters in configuration $C2$, i.e., (*0.01*,*5e-4*) can be assumed to be the best among all the different configurations considered for training the *SMSNet*, and the values of η and γ corresponding to this configuration have been used to report the results for the following experiments.

Next, we test the stability of the proposed network *(SMSNet)* by studying if the model performs with a similar level of effectiveness on the same test set even if it is initialized differently and trained with different training sets five times. The same set of model hyper-parameters determined in the previous experiment has been used here as well. The *CUHK_01*, *CUHK_03*, *Market1501*, and *VIPeR* data sets have been used to report

the results for this experiment. 90% samples are selected randomly from the training and validation samples of each data set (refer to Table 3.2) five different times and accordingly five different data sets are created, using which we train the *SMSNet* five different times from scratch. On completion of the training each time, we evaluate the performance of the trained model on the test set which is kept fixed. Figure 3.5 presents the results of this experiment in terms of the box and whiskers plot. The four boxes

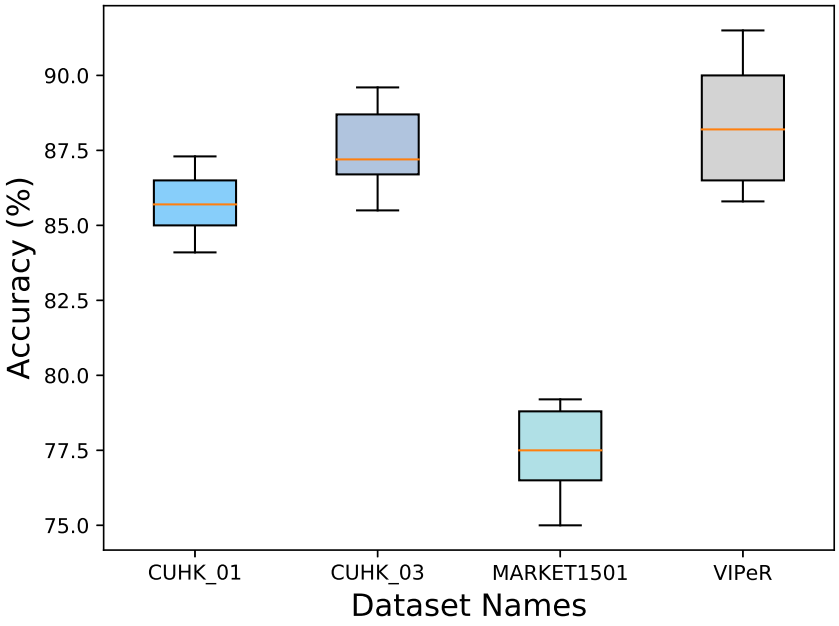

Figure 3.5: Box and whiskers plot showing the performance of the proposed approach after five different times of run on four data sets i.e., *VIPeR*, *CUHK_01*, *CUHK_03*, and *Market1501*

in the figure correspond to the range of accuracy obtained for each of the data sets used in the study, namely, *CUHK_01*, *CUHK_03*, *Market1501*, and *VIPeR* from the five runs. With reference to the figure, it can be observed that the inter-quartile range (i.e., between 25^{th} to 75^{th} percentile) corresponding to the *CUHK_01*, *CUHK_03*, and *Market1501* data sets are 1.5%, 1.7%, and 2.6%, respectively, which are quite small. The corresponding number for the *VIPeR* data set is 5% which is slightly larger than

the others, and this is because the network does not get trained properly due to the availability of limited training data. The small range of the whiskers in Figure 3.5 emphasizes the robustness of our approach against a wide variety of data sets.

We next compare our proposed approach with other popular state-of-the-art Siamese network-based techniques developed for image-based re-identification, namely, [130], [131], [1], [2], along with two non-Siamese network-based techniques, namely, DeepReid [3], and MuDeep [61]. Results are shown in Table 3.3 in terms of Rank 1 accuracy percentage. For this experiment also, we use a similar training-test set combination as already discussed in Section 3.1.3. The first two rows in Table 3.3 correspond to the two Non-Siamese network-based approaches while the rest of the rows show the performance of the Siamese network-based approaches, as mentioned above. From Table 3.3, it

Table 3.3: Comparison of Rank 1 accuracy (in %) for 100 test_ids of our proposed approach with state-of-the-art techniques

Methods	Rank 1 Accuracy (%)			
	VIPeR	***CUHK_01***	***CUHK_03***	***Market 1501***
Non-Siamese based				
Li et al. [3]	56.1	27.9	20.6	44.4
Qian et al. [61]	44.7	79.6	82.4	71.2
Siamese based				
Ahmed et al. [1]	35.2	64.2	55.0	56.7
Subramaniam et al. [2]	68.7	81.2	72.3	76.7
Varior et al. [130]	68.7	-	57.3	61.6
Guo et al. [131]	50.9	**88.1**	88.3	-
Proposed *SMSNet*	**91.5**	87.3	**89.6**	**79.2**

can be observed that the proposed *SMSNet* model for person re-identification usually performs better than the state-of-the-art approaches in terms of accuracy. Only in the case of the *CUHK_01* data, our approach falls short of the accuracy obtained from [131] by a very small percentage of 0.8. However, in all other situations, our approach stands out to be the winner. The superior performance of the proposed *SMSNet* on the *VIPeR* data set is since unlike the existing approaches, here our model is first trained on an

extensive data, namely, *CUHK_03* data, and next fine-tuned using the *VIPeR* data. This prevents the network from getting under-fitted due to the presence of a small number of training examples, thereby improving its accuracy.

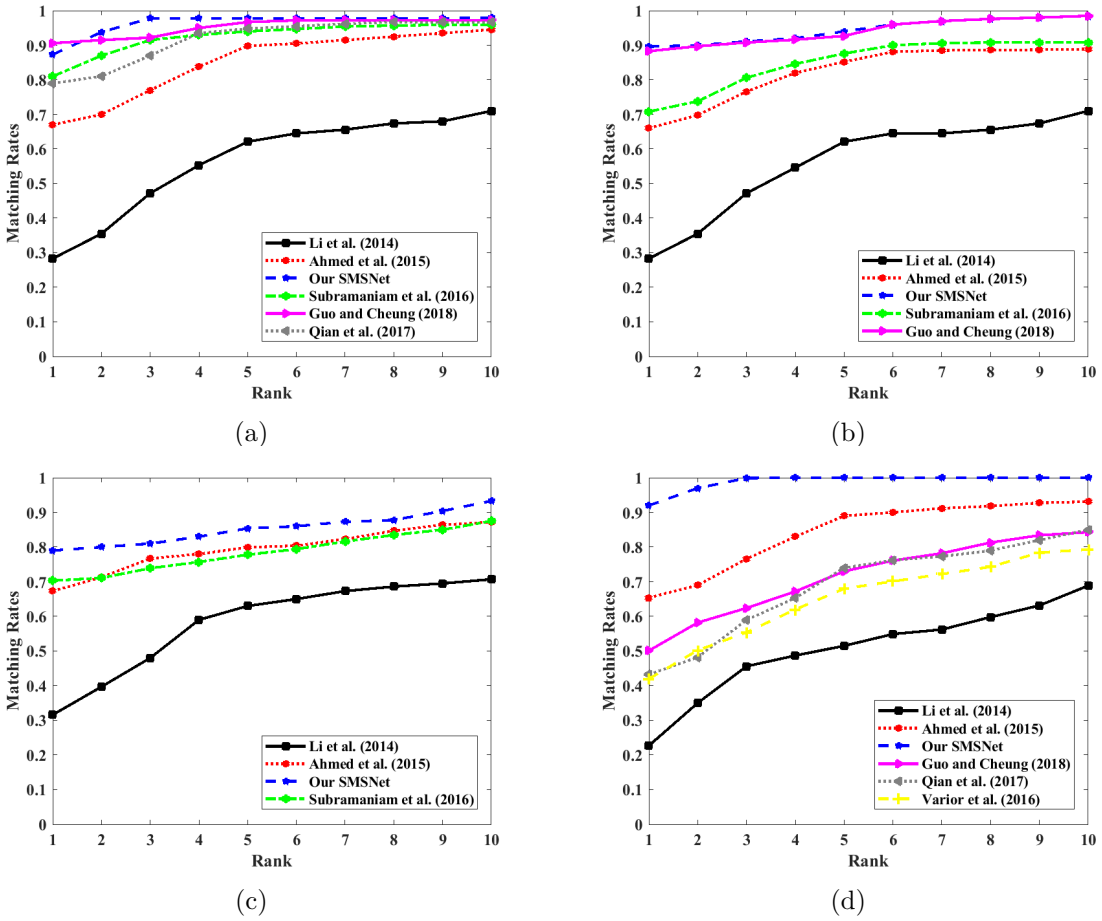

Figure 3.6: Cumulative matching characteristic curves showing improvement in re-identification accuracy with rank for the different approaches corresponding to: (a) *CUHK_01*, (b) *CUHK_03*, (c) *Market1501*, and (d) *VIPeR* data sets

The CMC curves corresponding to the different data sets used in the study, namely, *CUHK_01*, *CUHK_03*, *Market1501*, and *VIPeR* data are also presented in Figures 3.6 (a)-(d) respectively for Rank 1 to Rank 10. Once again, it is observed from the CMC curves that our proposed model provides high accuracy for most rank values and also for the different data sets. Although the Rank 1 accuracy of our approach on *CUHK_01* data is lower than that of [131] (as seen in Table 3.3), from Rank 2 onwards,

our approach performs better than [131] throughout. In general, the rank-wise accuracy of each of the other competing techniques is considerably lower than our approach for the different rank values. Also, it is observed that our method achieves a 90% or higher accuracy mark at Rank 1 for the *VIPeR* data, and at Rank 2 for both the *CUHK_01* and *CUHK_03* data, and at Rank 7 for the Market1501 data. We also observe that the average Rank 5 accuracy of our work is 96.02%, which is better than that of [131] (i.e., the approach with the second-best performance (85.65%)) by about 10%, which is remarkable.

3.1.4 Limitations of the Approach

Although use of dilation in the convolutional layers helps in capturing better low-level information from the images by increasing the receptive field of the filters, a limitation of this approach is that it makes the network more complex by incrementing the number of trainable parameters. Also, there is a possibility of losing intrinsic high-level structural details such as edges and textures. Understandably, eliminating the dilation in the convolutional layers may help in simplifying the *SMSNet* architecture considerably, but the effectiveness of the extracted features may suffer. To overcome this and achieve a similar level of effectiveness as that of the *SMSNet*, a hierarchical classification scheme may be employed in which color-based matching is used as an initial step to shortlist the best matches from the gallery set, and next re-identification is carried out using Siamese network from the reduced gallery set. Such an approach would enable the final re-identification to be performed by comparing the target image with only the closely matched samples and prevent the search from getting biased towards a completely different element with similar structural features. We discuss this proposed hierarchical classification approach in detail in the following section, i.e., Section 3.2.

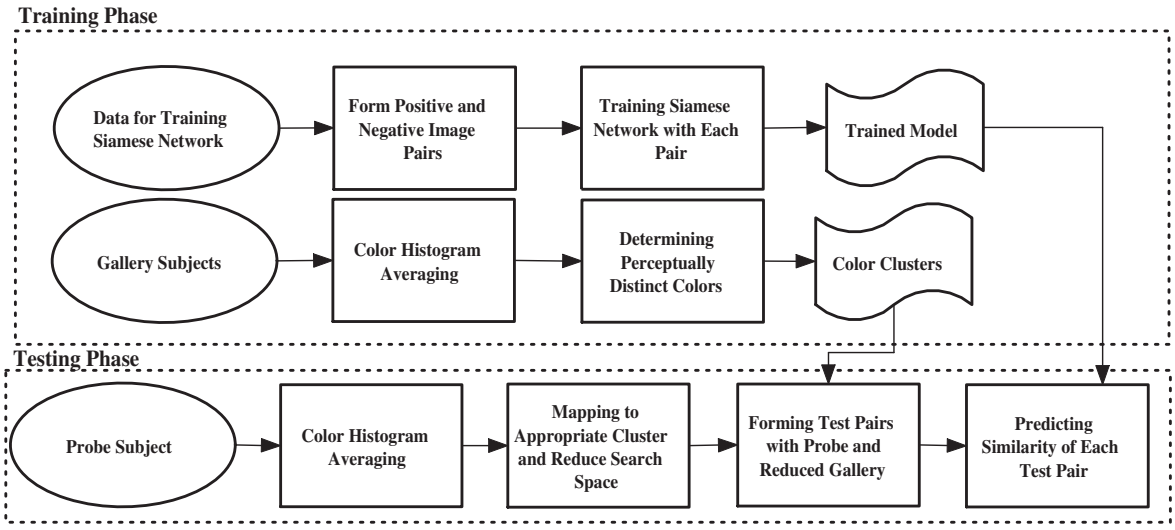

Figure 3.7: A block diagram of the proposed hierarchical approach to person re-identification

3.2 Hierarchical Classification for Person Re-Identification

A block diagram of the proposed approach is shown in Figure 3.7. With reference to the block diagram, the first phase of the approach is to train the Siamese network using positive and negative image pairs. The color histogram-based features are also computed from the gallery subjects and these features are clustered into the optimal number of clusters using K-Means clustering algorithm. During deployment, the color histogram features computed from the input test subject are first compared with all the cluster centers using Euclidean distance. Only those samples for which the corresponding cluster centers match closely with the features of the test image are considered for further comparison through Deep Siamese-based features.

In contrast to the previous approach discussed in Section 3.1, here we use a more simplified Siamese network with lesser trainable parameters and term it as the *Siamese Convolution Box (SCB)*. A detailed description of the *SCB* network architecture is given in Section 3.2.1, following which the proposed hierarchical classification scheme is described in Section 3.2.2. Finally, an extensive experimental evaluation is presented in Section 3.2.3.

3.2.1 *SCB* Training

The *SCB* network used in the present work is a four-layer network in which two layers are tied-convolved and the rest two are normal convolution layers. The complete architecture of the network is shown in Table 3.4, and the re-identification framework is shown in Figure 3.8. The framework is similar to that shown in Figure 3.3, the only

Table 3.4: Layer specification of the *SCB* network. Both the starting convolution layers are pooled with size 2×2 and the last two layers are exempted from pooling. ⋆ represents the concatenation of fully connected layers

Network	Layer	Size of filter	No. of filters
Siamese Convolution Box (*SCB*)	Conv2d_1	5×5	20
	Conv2d_2	5×5	25
	Conv2d_3	5×5	25
	Conv2d_4	3×3	25
Fully Connected	**Layer**	**No. of neurons**	
	fc_1	500	-
	fc_2	500	-
	fc_3	500	-
	FC	fc_1⋆fc_2⋆fc_3	-

difference being here the *SMSNet* block has been replaced with the *SCB* block, and the final step of re-identification has been carried out using only a small gallery set. A magnified view of the *SCB* is shown in Figure 3.9. The figure shows that after the first two convolution layers, there is a feature difference layer that subtracts the features obtained from the first two layers. This difference feature is further distilled using two more convolution layers without pooling to prevent significant information loss due to shrinking after the first two layers.

As seen in Figure 3.8, three different Siamese blocks are trained on the different silhouette parts, namely, the head, torso, and leg parts, by dividing the silhouette equally into three regions. The latent vectors obtained from each block (namely, fc_1, fc_2, and fc_3) are compressed to obtain the final fully connected layer (*FC*). The network is next trained on an extensive data set constructed from the *CUHK_03* [3] and *Market1501* [108] re-identification data sets. The *SCB* is trained in multiple epochs using Adam optimizer until convergence and at each epoch during training the soft-

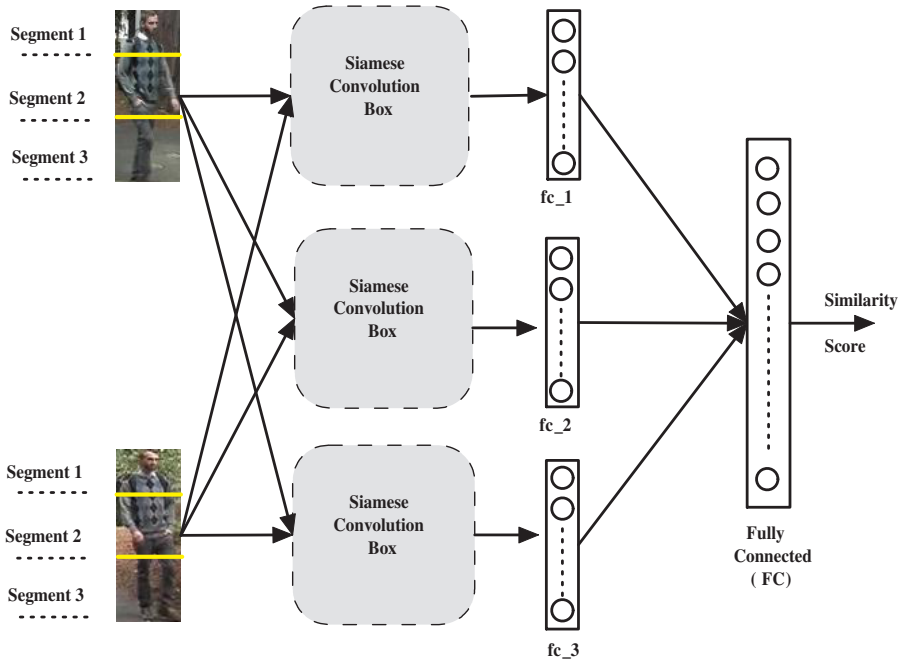

Figure 3.8: Re-Identification framework applied on the reduced gallery set

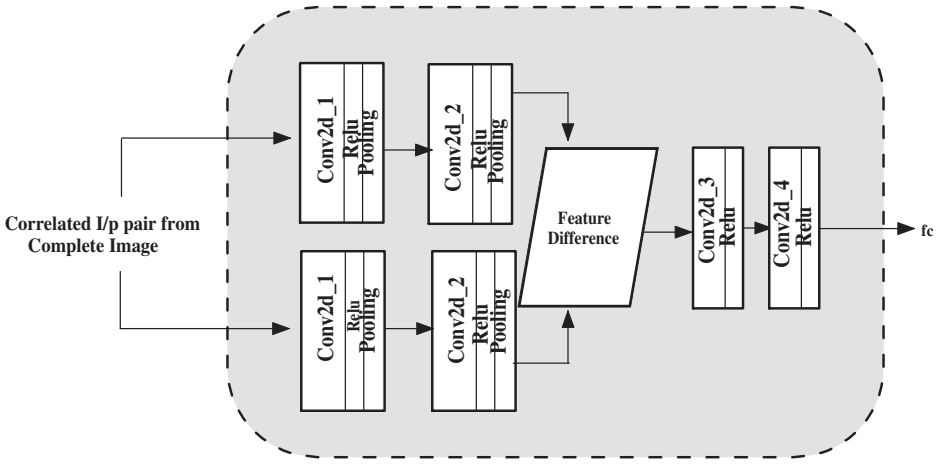

Figure 3.9: Siamese Convolution Box (*SCB*)

max cross-entropy function is used at the final layer of the network to measure the network loss. Convergence is said to have reached if the difference in the network losses between two successive epochs reaches a pre-determined small threshold value ϵ. In our experiments, the value of ϵ is considered as 10^{-5}.

3.2.2 Hierarchical Approach to Re-Identification

As a first step, we aim to group all the gallery subjects into separate clusters based on similar color appearance features. Color histograms in the R, G, and B channels are used as features to perform this grouping. For each frame, first background subtraction, RGB silhouette extraction, and normalization to a fixed height and width need to be carried out using standard techniques [1], [66]. Since different silhouettes with similar color distribution are likely to have the same histogram pattern, we first segregate each silhouette into three equal parts as shown in Figure 3.8, and carry out silhouette part-based analysis by computing the histogram for each of the three parts in the R, G, and B channels. Each of these channels is further quantized into 16 bins to eliminate the effect of appearance changes due to illumination differences in the two camera views. Next, color histograms computed from all the frames of a walking sequence are averaged to generate the final color appearance descriptor H^i of the i^{th} subject. The procedure used to obtain the number of clusters is briefly discussed next.

Assuming the gallery set contains N subjects, the histograms computed from all the N subjects (i.e., H^1, H^2, ..., H^N) are concatenated to form a feature matrix H, which is next clustered into a fixed number of groups using K-means clustering. The clustering is done such that subjects with similar perceptual color appearances (i.e., similar H^i features) are placed in the same cluster. The number of color groups (i.e., K) to be formed from the set H is determined by plotting an elbow curve. This curve shows the variation of the clustering error (i.e., the summation of the square of the intra-cluster distances) as the value of K is gradually increased from a small value. If $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, ..., $\mathcal{C}_K$ are the K cluster centers at a given point of time, then the clustering error is computed as:

$$E = \sum_{i=1}^{N} ||H^i - C_{H^i}||^2, \tag{3.2}$$

where C_{H^i} denotes the cluster center to which H^i gets mapped, $||.||^2$ denotes the Eu-

clidean norm, and $1 \leq H^i \leq$K. The elbow curves obtained from the *VIPeR* and the *CUHK_03* data sets are shown in Figures 3.11(a) and (b), respectively. From each of these figures, it can be seen that the clustering error does not reduce significantly as the number of clusters is increased beyond 100. Hence, a value of K equal to 100 can be considered to be an optimal choice from the elbow curve. Similar approaches to determine the optimal number of clusters can be found in different applications in the past, e.g., [132, 133].

As soon as a particular subject S'_t appears in the field of view of the camera Cam_2 positioned above the exit point, the averaged color histogram (say, H'_t) of the test subject is computed and the top $\mathcal{K}$ matching clusters are selected. For example, if $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_\mathcal{K}$ are the top $\mathcal{K}$ matching clusters corresponding to the subject S'_t, then the subsequent Siamese network-based comparison is done on subset S_{red} of S such that $S_{red} \in \{S_{C_1} \bigcup S_{C_2} \bigcup ... \bigcup S_{C_\mathcal{K}}\}$, where S_{C_k} is the set of subjects belonging to cluster k (k = 1,2,3,...,$\mathcal{K}$). Reducing the search space by eliminating dissimilar elements helps in improving the re-identification accuracy by preventing the subsequent Siamese network-based prediction stage from getting biased towards an incorrect element in the gallery set having similar structural features.

Finally, the test subject S'_t is compared with all the subjects in the reduced set S_{red} determined from the previous cluster matching stage. Since both Cam_1 and Cam_2 capture videos instead of still images, the averaged normalized silhouettes computed from the two video sequences are provided as input to the *SCB* network. The normalized silhouettes are obtained by finding the bounding box around the target subject and resizing it to a fixed height and width. The silhouette normalization step has already been done in the data sets used in the study, and hence we did not have to carry out this step in the thesis. The test subject S'_t is assigned the class $\mathcal{T}$ if

$$sim(\mathcal{T}, S'_t) > sim(X, S'_t), \forall X \in S_{red}, X \neq \mathcal{T}, \tag{3.3}$$

where $sim(A, B)$ represents the similarity score given by the Siamese network for two input images A and B.

3.2.3 Experimental Evaluation

As for the previously proposed approach given in Section 3.1, here also for evaluation, we consider the same four data sets, namely, *VIPeR* [106], *CUHK_01* [107], *CUHK_03* [3], and *Market1501* [108]. Additionally, to simulate the deployment scenario shown in Figure 3.1, we capture a new data set in the laboratory that consists of walking videos of 41 subjects and evaluate our approach using this data set as well. This data set is termed as the *IIT (BHU) Re-Identification data* and a detailed description of the data set has already been given in Section 2.4 of Chapter 2.

The performance evaluation of the proposed *SCB* has been done using the same training-test combination as explained earlier in Section 3.1.3. The two user-defined constant parameters to be specified before training the *SCB* network are the learning rate (η) and the weight decay factor (γ). To get a good estimate of these parameters, we study the average cross-validation accuracy for different combinations of the above parameters and select the one that provides the maximum cross-validation accuracy. Specifically, we consider three different pairs of η and γ values as follows: $C1$(*0.01*,*0.0250*), $C2$(*0.01*,*0.0005*), and $C3$(*0.07*,*0.0005*), and train the *SCB* from scratch fives times for each of these three configurations. The one with the best average cross-validation accuracy is finally selected and has been used to report the results for all future experiments. For a particular combination of the η and γ values, the cross-validation scheme is carried out by constructing three different pairs of training and validation sets by randomly partitioning the gallery set, and next computing the average cross-validation accuracy obtained from these three validation sets. Since, we perform a *leave-100-out-cross-validation*, in each of the above partitions the cardinality of the validation set is considered to be 100. The average cross-validation accuracy for the above-mentioned

three data sets and the different combinations of η and γ values are reported in form of box plot in Figure 3.10. With reference to the figure, the horizontal axis represents the

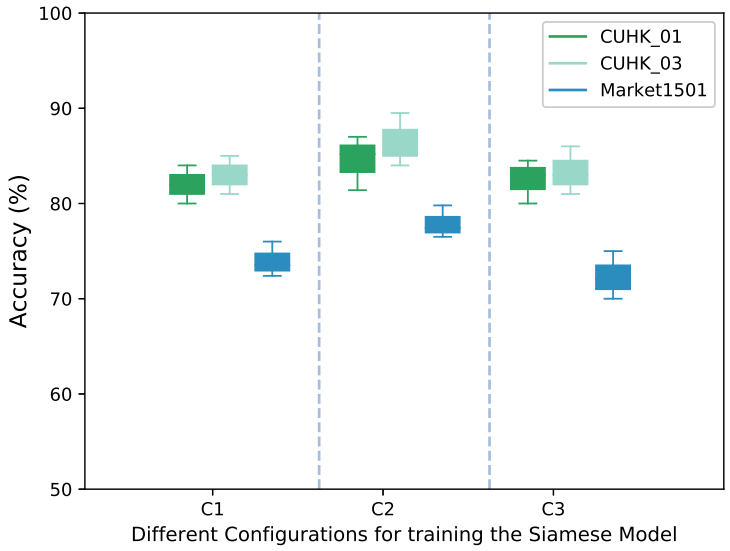

Figure 3.10: Range of five-fold cross-validation accuracy for various combinations of parameters η and γ corresponding to different the data sets by setting different initial weights of the network

different sets of hyper-parameters (i.e., η and γ) whereas the vertical axis represents the Rank 1 accuracy. We have observed that the maximum cross-validation accuracy corresponding to each data set has been obtained for η=0.01 and γ=0.0005. Hence, these values for η and γ have been used to train the Siamese network *(SCB)* in all future experiments. Our observation is that the training algorithm converges as the number of epochs reaches 80000 for each of the *CUHK_01* and *CUHK_03* data sets, while for the *Market1501* data the corresponding number is 110000.

In our next experiment, we determine the optimal number of perceptually distinct colors from a given data set to perform re-identification. For this, we study the improvement in the clustering error with increment in the number of clusters (K) for the *VIPeR* and the *CUHK_03* gallery sets. The respective plots are shown using elbow curves in Figures 3.11(a) and (b), and in each of these figures the horizontal axis represents the number

of clusters while the vertical axis denotes the clustering error for the corresponding number of clusters.

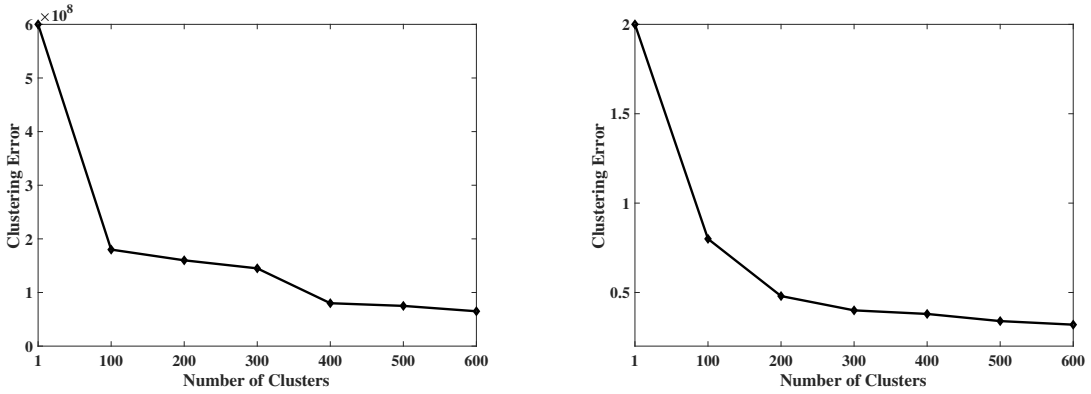

Figure 3.11: Elbow Curves for (a) *VIPeR* and (b) *CUHK_03* data

From both the figures, it is observed that the clustering error is quite low if the number of clusters is set to be equal to 100, and this is the point in the plots where the first elbow can be observed. It is also seen that beyond 200 clusters, the curve attains almost a saturation level. Thus, the optimal number of color clusters for the *VIPeR* and *CUHK_03* data sets may be assumed to lie between 100 and 200. Similar elbow curves obtained from each of the other data sets (except the *IIT(BHU) Re-Identification Data Set*) reveal that the optimal number of clusters falls in the range [100,200], while in the case of the *IIT(BHU) Re-Identification Data Set*, the corresponding number lies within the range [5,10]. To avoid the repetition of similar results, these plots have not been presented in the chapter.

Next, to obtain a good estimate of the parameters K and $\mathcal{K}$ (refer to Section 3.2.2), we study the effect of varying these parameters on the average cross-validation accuracy on the five public data sets, namely, the *VIPeR* [106], *CUHK_01* [107], *CUHK_03* [3], *Market1501* [108]. Table 3.5 shows this average cross-validation accuracy on five different validation sets derived from the gallery set of each data set as the value of K is increased from 100 to 500 in steps of 100, and for two values of the parameter $\mathcal{K}$,

namely, *1* and *2*. The average response times to re-identify a subject for the different experimental settings have also been reported in the same table. It is verified from

Table 3.5: Average cross-validation accuracy of the proposed approach on different data sets and average response time (in milli-secs)

Data set	Performance Evaluation	Selecting Clusters based on Color Matching							
		$\mathcal{K}$=1				$\mathcal{K}$=2			
	Clusters	100	200	300	500	100	200	300	500
VIPeR	**Acc**(%)	88.3	85.4	78.6	76.5	88.0	86.4	77.5	73.8
	Time(ms)	40.5	31.2	25.5	13.6	43.8	32.4	27.6	14.1
CUHK_01	**Acc**(%)	82.1	83.4	76.8	72.2	84.2	85.1	78.6	75.2
	Time(ms)	43.0	26.3	14.2	9.5	45.2	32.6	18.5	12.2
CUHK_03	**Acc**(%)	91.0	89.6	83.1	81.6	92.7	90.0	82.6	81.1
	Time(ms)	41.2	29.1	15.2	10.3	46.5	32.4	19.9	11.6
Market 1501	**Acc**(%)	93.6	89.4	84.8	81.0	93.2	90.7	85.6	80.3
	Time(ms)	49.9	37.8	25.0	14.3	52.7	40.0	25.6	16.8

the table that, as expected, higher cross-validation accuracy values are obtained if the value of K is set to be equal to 100 or 200. In general, this accuracy decreases as the value of K is increased, and this reduction in the re-identification accuracy is due to the elimination of clusters with the correct matches during the cluster mapping phase. From the above results, it is clear that the clustering step based on the color histogram features (as explained in Section 3.2.2) has a profound influence on the final cross-validation accuracy. To study the impact of the clustering phase on the final prediction accuracy, in the next experiment we observe the variation of the accuracy of the cluster mapping phase with increment in the number of clusters. The training set constructed from each data set is used to determine the color clusters, following which each sample from the validation set is mapped to the appropriate cluster, and finally, the cluster mapping accuracy is computed. A validation sample is said to map to the correct cluster only if the training sample of the same subject also gets mapped to the same cluster. From the knowledge of the ground truth, the accuracy of this cluster mapping phase can be computed. Similar accuracy values are computed by varying K from 50 to 100 in steps of 50. A plot of cluster mapping accuracy with respect to the number

of clusters for the *CUHK_03* data is shown in Figure 3.12.

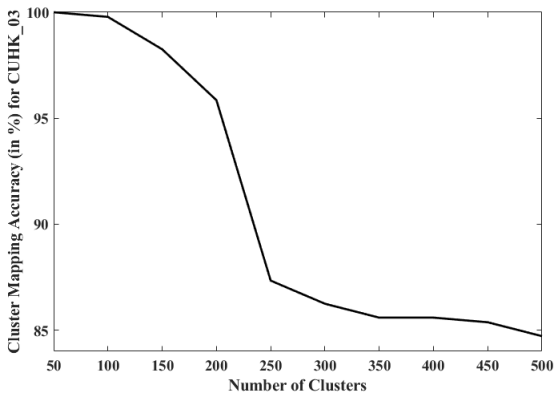

Figure 3.12: Variation of the cluster mapping accuracy with increment in the number of clusters for *CUHK_03* data

It can be seen from the figure that while for 50 and 100 number of clusters, the cluster mapping accuracy is close to 100%, it reduces as the number of clusters continues to increase, and attains about 85% accuracy mark when the number of clusters reaches 500. A similar observation has also been obtained for each of the other data sets as well, but these plots have not been presented in the chapter to avoid repetition of similar results. From the results of Table 3.5 and Figure 3.12, it can be concluded that while the clustering phase can significantly improve the effectiveness of the re-identification algorithm by reducing the search space for identity matching, finding an optimum partitioning from the gallery samples is crucial to carry out re-identification accurately. If a high value for K is chosen for clustering, it might be possible that during the cluster building phase, multiple clusters get formed out of subjects with similar appearances (i.e., color histogram features). This, in turn, increases the chance for a test sample to get mapped to a similar but incorrect color cluster during the cluster mapping phase. As explained in Section 3.2.2, $\mathcal{K}$ is a parameter that decides how many top matching clusters should be retained for the second level of classification using the Siamese network. As expected, Table 3.5 shows that the response time increases if a higher value of $\mathcal{K}$ is chosen keeping other factors constant. It can also be seen from

the table that $\mathcal{K}$=2 performs marginally better than $\mathcal{K}$=1, which is logical since the search space for re-identification increases for higher values of $\mathcal{K}$. Although setting the value of $\mathcal{K}$ to *2* provides marginally better results, it reduces the efficiency of the re-identification process. On the other hand, by setting the value of $\mathcal{K}$ to *1*, more than 82% accuracy is obtained for each of the five data sets corresponding to 100 clusters, within a considerably shorter time. Hence, *K=100* and $\mathcal{K}$*=1* can be considered to be optimal parameters for re-identification for each of the above data sets. It may be noted that the optimal values for the parameters K and $\mathcal{K}$ are data-specific. For a different data set, another set of parameter values might turn out to be optimal. Hence, given any re-identification data set, the optimal values for K and $\mathcal{K}$ have to be first determined from the gallery set before carrying out re-identification using the test samples.

Each of the above experiments deals with parameter tuning and model training using the gallery set data consisting of training and validation samples. Next, we evaluate the effectiveness of the proposed approach on unseen data over other popular state-of-the-art techniques using the test samples corresponding to each data set. We perform a comparative performance analysis of our work with the same approaches used for experimental evaluation in Section 3.1.3, i.e., [1–3, 61, 130, 131]. The trained model for each data set, as obtained from the previous experiments, has been used to report the results of this experiment on the test set of 100 samples. In this experiment, we study the robustness of the different re-identification approaches by observing the variation of Rank 1 re-identification accuracy values for different non-overlapping subsets of the test set derived from each data set. Each test set of 100 subjects is divided into four non-overlapping subsets of 25 subjects, following which the accuracy of the trained models on these different sub-sets are observed. Let us denote these four test sets as *Test 1*, *Test 2*, *Test 3*, and *Test 4*. Figures 3.13(a)-(e) respectively present the Rank 1 re-identification accuracy of the above-mentioned approaches on the different test sets in the form of grouped bar charts corresponding to the six datasets used in the

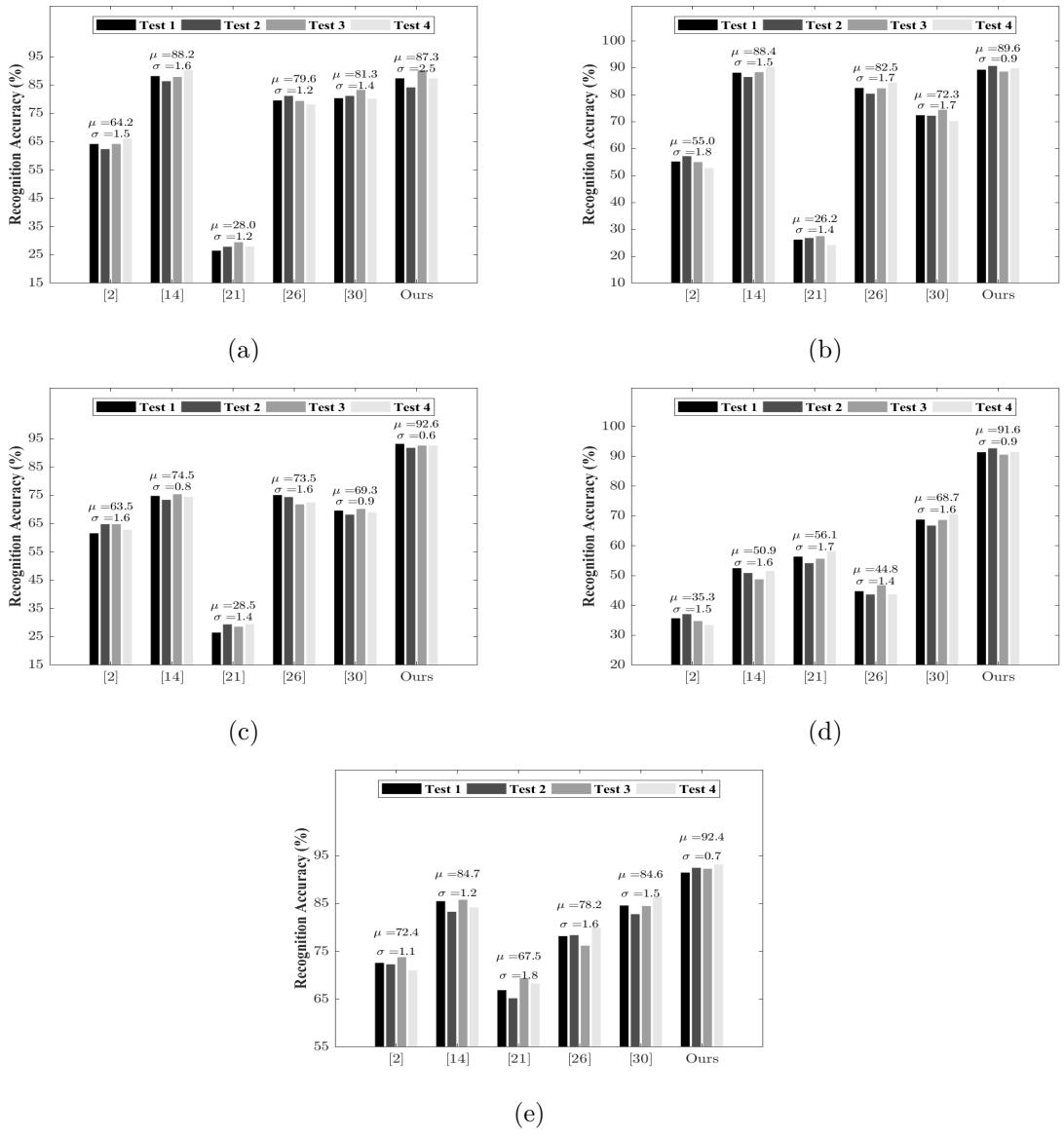

Figure 3.13: Accuracy and standard deviation of accuracy values on the four test sets constructed from (a) *CUHK_01*, (b) *CUHK_03*, (c) *Market1501*, (d) *VIPeR*, and (e) *IIT(BHU) Re-Identification* Data

study, namely, the *CUHK_01*, *CUHK_03*, *Market1501*, *VIPeR*, and *IIT(BHU) Re-Identification* data. In each plot, the horizontal axis shows the citations of the different re-identification methods, while the height of each bar shown along the vertical axis represents the corresponding re-identification accuracy. The μ and σ values reported on the top of each group of bars represent the mean and the standard deviation of the observations, respectively. From the figures, it can be seen that, in general, our method performs better than the state-of-the-art approaches both in terms of re-identification accuracy and response time. Only in the case of the *CUHK_01*, and *CUHK_03* data sets, the average performance of our approach is closely comparable to that of [131]. While for the *CUHK_01* data, the average accuracy of the work of [131] is about 0.8% better than that of ours, for the *CUHK_03* data our approach surpasses the accuracy of [131] by about 1.3%. However, for each of the other data sets the average performance of our method is significantly better (more than 12%) than each of the other existing re-identification techniques used in the study. This superiority in the performance is mostly due to the hierarchical matching scheme followed in this work that eliminates vastly dissimilar candidates after the first stage, thereby preventing the second stage of Siamese network-based classification from getting biased towards a different element in the gallery set with similar structural features. Also, the low standard deviation value of our approach in each figure emphasizes the fact that our approach performs consistently well on each of the different test sets, namely, the *Test 1*, *Test 2*, *Test 3*, *Test 4*. We have also noted the average response times for the different approaches to compare between a pair of gallery and test subjects. These are *0.09*, *0.10*, *0.16*, *0.08*, *0.11*, and *0.06 seconds*, respectively for the approaches $[1, 2, 130, 131]$ and our work. Hence, it can be concluded that the proposed approach outperforms state-of-the-art re-identification techniques both in terms of accuracy and efficiency.

We next perform a comparative rank-based performance analysis (from Rank 1 to Rank 10) of our method with other state-of-the-art re-identification approaches using Cumu-

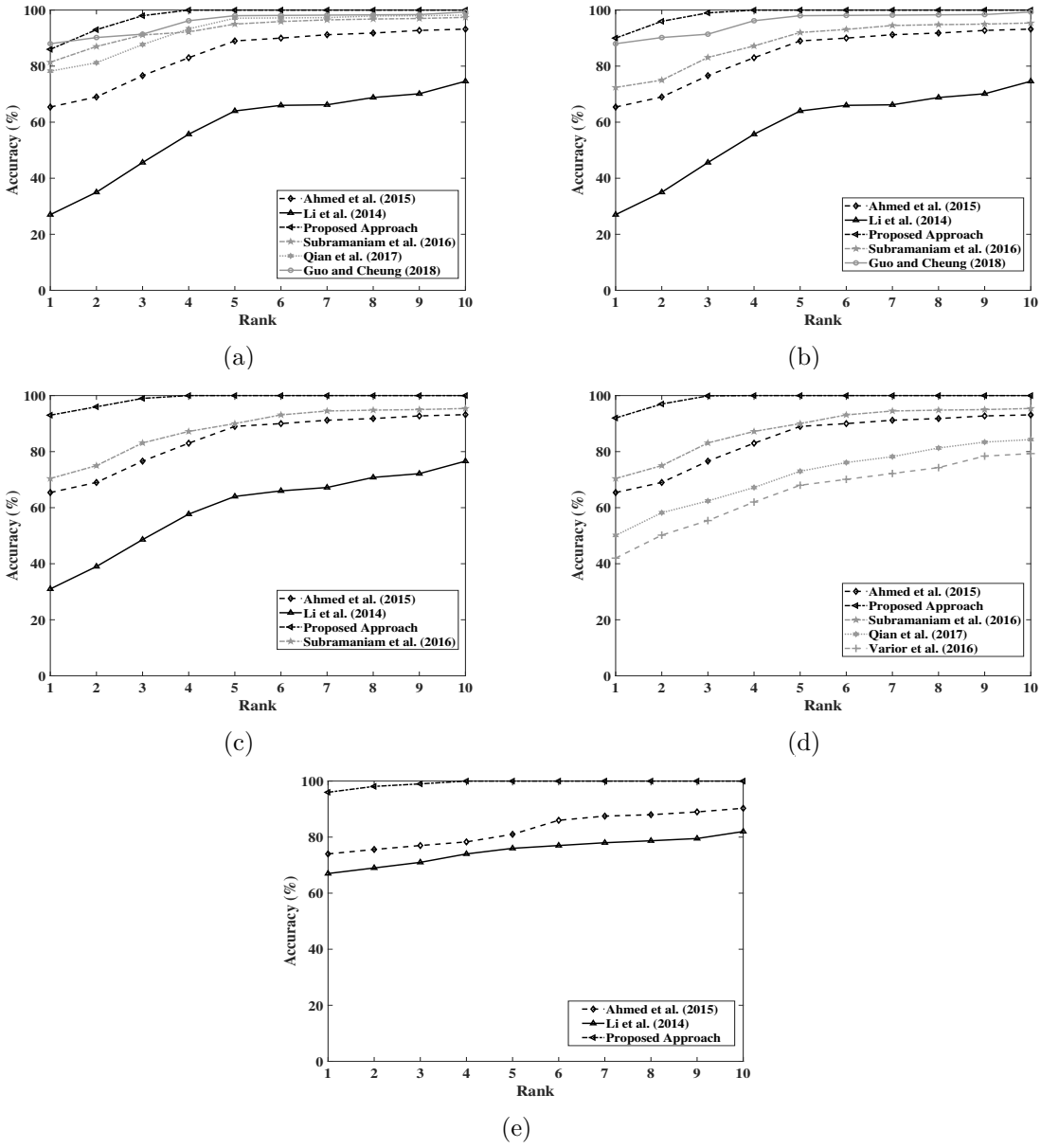

Figure 3.14: Comparative performance analysis of the different re-identification approaches by means of Cumulative Matching Characteristic curves on (a) *CUHK_01*, (b) *CUHK_03*, (c) *Market1501*, (d) *VIPeR*, and (e)*IIT(BHU) Re-Identification* Data

lative Matching Curves (CMC) on the complete test set of 100 ids. Figures 3.14(a)-(f) respectively present the CMC curves obtained for the different re-identification approaches (namely, [1–3, 61, 130, 131]) on the six data sets used in the study. The rank-based re-identification accuracy obtained for the data sets *CUHK_01* [107], *CUHK_03* [3] and *Market1501* [108] are presented in Figures 3.14 (a), (b), and (c), respectively. We also perform cross-data set experiments and plot the rank-based improvement in re-identification accuracy of the different techniques on the *VIPeR* [106], and *IIT(BHU) Re-Identification* data in Figures 3.14 (d) and (e), respectively. From the figures, it can be again observed that the proposed approach always outperforms the state-of-the-art re-identification methods by a large margin. Moreover, our approach achieves the 100% accuracy mark on the test set of all the data sets only within a rank of 4. As already explained from the results of Figures 3.13(a)-(e), the improved performance of the proposed technique is mostly due to the use of tied-convolved layers in the *SCB* and employment of the hierarchical classification scheme that retains only the top matches from the gallery set to make the final prediction.

As explained in Section 3.2.2, the proposed re-identification algorithm consists of two major components: (i) determining the appropriate cluster/s by utilizing the color information of the test subject, thereby reducing the search space, (ii) comparing the test subject with the elements of the reduced set using a Siamese network. In the following experiment, we study the effect of these individual components on the overall accuracy and response time. Specifically, we carry out a rank-based performance analysis of our proposed work with two different settings using the test sets of the *CUHK_01* [107], *CUHK_03* [3], and *Market1501* [108] data sets. These are (i) re-identification by incorporating the clustering phase using the optimal values for K and $\mathcal{K}$ as determined during training, and (ii) re-identification without including the cluster matching component. The same training-test pair, as well as the trained Siamese network determined in the training phase, have also been used to report these results. Figures 3.15 (a), (b), (c)

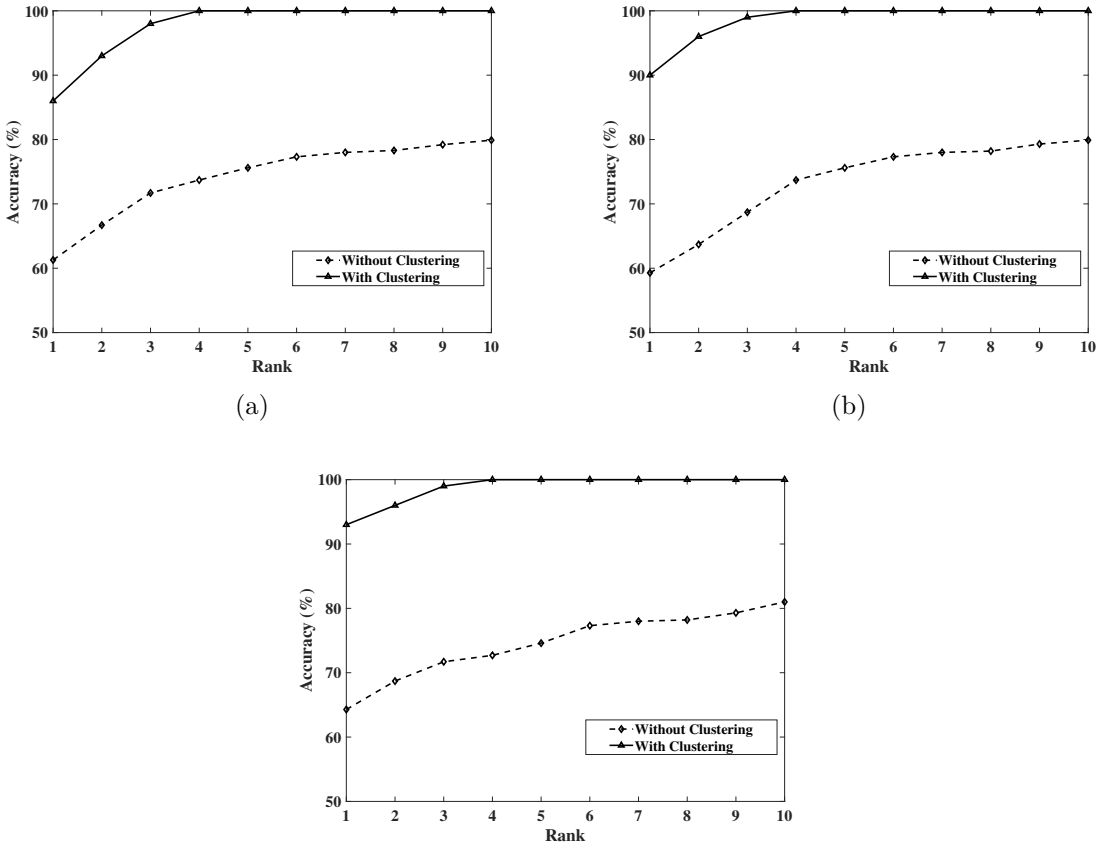

Figure 3.15: Cumulative Matching Characteristic curves showing improvement in re-identification accuracy with the rank of the proposed algorithm with and without the clustering phase for the (a) *CUHK_01*, (b) *CUHK_03*, (c) *Market1501* data sets

show the CMC curves representing the rank-wise improvement of the re-identification accuracy corresponding to the above three data sets for each experimental setting. It is seen from the figure that the clustering step (i.e., the first level of the proposed hierarchical re-identification method) has a significantly high impact on the final re-identification accuracy. In each of the curves, the re-identification accuracy for the different ranks can be seen to be much higher when the clustering phase is employed. Also, the average times to re-identify a test subject with and without clustering are 0.454 secs and 0.232 secs, respectively, and hence it can be assertively said that the inclusion of the clustering phase does not reduce the efficiency of the re-identification

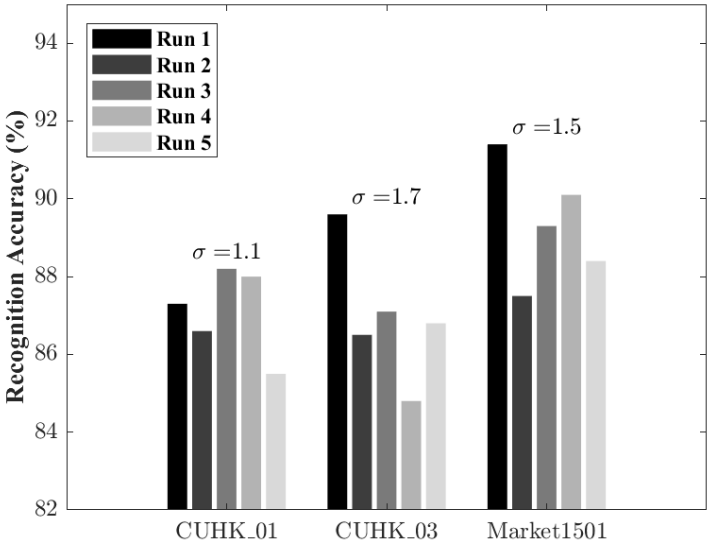

Figure 3.16: Percentage accuracy and the standard deviation of the accuracy values on 100 test samples of the *CUHK_01*, *CUHK_03*, and *Market1501* data sets obtained by training our model with different initialization parameters five times

process much.

The above experiments verify the robustness of the proposed approach for the different training-test set combinations. Further, to evaluate the reliability of our approach, we study the re-identification accuracy given by our approach for five different random initialization of the weights of the Siamese network and training the network via cross-validation using the gallery set. The *CUHK_01*, *CUHK_03*, and *Market1501* data sets have been used to perform this experiment. Each random initialization results in a differently trained model, and the performance of each of these five trained models is next evaluated on the test set through Rank 1 re-identification accuracy. The corresponding results are plotted using grouped bar charts in Figure 3.16. Here, the horizontal axis refers to the data set names, while the vertical axis refers to the re-identification accuracy. The standard deviation (σ) of the accuracy values obtained after running the five differently trained models on the same test set is shown on the top of each group of bars in the same figure. It is observed from the results that, for the above-mentioned

three data sets, i.e., *CUHK_01*, *CUHK_03*, and *Market1501*, the prediction accuracy values vary in the range [81.3, 88.2], [84.8, 90.8], and [91.5, 91.4], respectively. Also, the standard deviations of the accuracy values obtained after five runs on these data sets are only 1.1, 1.7, 1.5, respectively, each of which is quite low. From the high average accuracy (above 85%) and low standard deviation of accuracy, it can be concluded that the proposed re-identification approach should be able to perform reliably for different training-test data combinations.

Since the first level of the hierarchical classification carries out short-listing based on color information, it appears that our approach might fail if the subjects in the re-identification gallery set wear similar colored clothes. To study the effectiveness of our approach in this challenging scenario, we study a rank-wise performance improvement of our approach with and without the cluster determination and mapping phases on a data set where subjects wear similar colored clothes. The subset of the *IIT (BHU) Re-Identification data set* constructed with 20 individuals (refer to Section 2.4.2 of Chapter 2) wearing similar colored clothes has been used for this experiment. The same pre-trained Siamese model used for presenting the results of Table 3.5, and Figures 3.14 and 3.15, has also been used here. The two CMC curves presented in Figure 3.17 represent the rank-wise performance improvement of the proposed algorithm with and without including the clustering phase on this data set. To obtain the results of the clustering phase, we determine the optimal values for K and $\mathcal{K}$ as *5* and *2*, respectively from the elbow curve in a similar manner as discussed in Section 3.2.2. It is observed from the figure that higher accuracy is obtained for each of Rank 1 and Rank 2 if the clustering phase is included. This is since the elimination of dissimilar candidates after the first hierarchical level improves the prediction of the Siamese network. However, for higher ranks, the approach without clustering performs better. This is due to the difficulty in determining the dominant colors from similar-looking individuals during the clustering phase. Often the correct match gets eliminated after the color matching phase, and

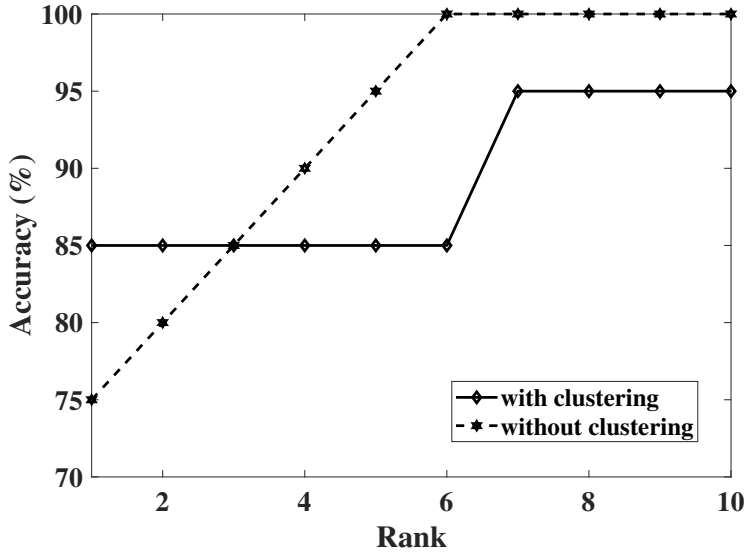

Figure 3.17: Cumulative Matching Characteristic curves showing improvement of re-identification accuracy with rank on a test set with similar clothing conditions with and without considering the cluster determination and mapping phases

the deep feature-based classification used in the second stage of the hierarchy has no chance of predicting the correct match in such cases. Thus, the proposed algorithm with the clustering phase never achieves the 100% mark even for higher values of the rank. This situation does not arise if the clustering stage is not included, since in this case the final prediction is obtained by comparing the test sample with the entire gallery set. However, even in this challenging scenario, the proposed approach (without the clustering phase) performs with 85% accuracy at Rank 1 and also achieves 100% accuracy within only Rank 6. This emphasizes the effectiveness of the proposed *SCB*-based feature extractor.

Handling images of persons with similar appearances has always been regarded as one of the challenging problems in person re-identification [134]. To evaluate the effectiveness of our approach over state-of-the-art techniques in handling person images with similar appearances, we next carry out a comparative performance evaluation of the different re-identification approaches, namely, [1–3, 61, 131] using the same training and test sets as used for the previous experiment. Results are shown in terms of Rank 1

accuracy in Table 3.6. The first five rows correspond to the re-identification accuracy of five state-of-the-art approaches, while the last two rows show the accuracy obtained using the proposed approach with and without incorporating the cluster-based matching stage. The results from the table reveal that the proposed approach (with clustering)

Table 3.6: Comparative study of Rank 1 accuracy of different re-identification approaches on a test set of subjects with similar clothing conditions

Methods	Accuracy (in %)
Ahmed et al. [1]	53.0
Guo et al. [131]	82.0
Li et al. [3]	41.0
Qian et al. [61]	81.0
Subramaniam et al. [2]	76.0
Proposed Approach	**85.0**
Proposed Approach (without clustering)	75.0

outperforms each of the state-of-the-art techniques by at least 3% even in the case of the challenging scenario where subjects wear similar colored clothes. This is due to capturing effective person-specific features through *SCB*, and also employing part-based silhouette analysis, as described in Section 3.2.1. From the extensive experimental evaluation described above, we conclude that the proposed re-identification approach outperforms the existing techniques used in the comparative study both in terms of accuracy and efficiency.

We next carry out a comparative performance analysis between the two approaches proposed in this chapter in Sections 3.1 and 3.2 using the same training and test set combination of *VIPeR*, *CUHK_01*, *CUHK_03*, and *Market1501* data sets, as discussed in Section 3.1.3. Table 3.7 presents the Rank 1 accuracy given by these two approaches on the different data sets. It can be seen from the table that the proposed hierarchical re-identification scheme based on *SCB* performs better than the one using *SMSNet* in terms of accuracy for each of the data sets. This is since *SMSNet* tends to lose certain edge and shape-related features due to putting a higher focus on the low-level information by increasing the receptive field of filters in the convolutional layers. In

Table 3.7: Comparison of Rank 1 accuracy (in %) for 100 test_ids

Methods	Rank 1 Accuracy (%)			
	VIPeR	***CUHK_01***	***CUHK_03***	***Market1501***
SMSNet	91.5	87.3	89.6	79.2
Proposed Approach	**93.0**	**88.2**	**90.8**	**91.2**

contrast, the tied convolution layers used in *SCB* share useful parameters across the layers which helps in coming up with a more accurate descriptor, thereby resulting in a better re-identification performance.

3.3 Summary

In this chapter, we have presented two effective methods of feature extraction through (i) a multi-scale deep learning framework and (ii) a hierarchical classification scheme, each of which employs Siamese network-based feature extractors by dividing a silhouette into multiple segments. In Section 3.1, we study the use of dilation in the Siamese network to derive effective multi-scale features. We introduce a Siamese model termed as *SMSNet* and apply different dilation rates in the convolution layers of the *SMSNet* to enable capturing of detailed visual features. Additionally, the silhouette part-based analysis presented in this work helps in preserving the spatial relationships among the different silhouette segments at a high resolution. Since the use of dilation in convolutional layers may lose important high-level image information and also increase the complexity of the network, we next study the use of a Siamese network without incorporating dilation in its layers. In this context, we propose a new model termed as the Siamese Convolution Box *(SCB)* and adopt a hierarchical classification scheme through an effective combination of a traditional passive-based approach and a modern Deep Learning-based approach. Here, color histogram-based matching is employed at the initial level of hierarchy to shortlist the top few closest matches and reduce the gallery set for the subsequent step of Siamese network-based matching. Incorporation

of the initial color-based matching scheme reduces the search space and prevents the Siamese network from getting biased towards an utterly different element with similar structural features that may be present in the gallery set. Results on several public data sets demonstrate that the SCB-based re-identification performs better than SMSNet-based re-identification and also outperforms the existing re-identification techniques by a high margin. It may also be noted that the re-identification methods discussed in this chapter do not need the person images to be captured strictly from the front view. Rather, these are equally effective in situations where the images are captured from near-front views or other views as well. However, the target subject must occupy the major portion of the cropped and normalized frame and must have a similar color-based appearance from each view.

As per the re-identification scenario discussed using Figure 3.1, both the entry and exit gate cameras capture human walking videos from the front view. However, each of the approaches discussed in this chapter computes only spatial-domain features from the image frames and does not exploit the temporal information present in the sequential video frames to derive the feature descriptors. While working with video data sets, we compute an averaged normalized frame from each sequence which is next used in the re-identification phase. Since we are dealing with surveillance scenarios, where cameras typically capture walking videos consisting of sequential frames, it appears that the use of temporal features derived from the motion information can play a vital role in improving the re-identification accuracy further, that we are going to study in the next chapter.