Person Re-Identification for Intelligent Surveillance Using Deep Learning

Thesis submitted in partial fulfillment for the Award of Degree

Doctor of Philosophy

by

Nirbhay Kumar Tagore

निर्भय कुमार टैगोर

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 17071509 *Year 2021*

Chapter 7

Conclusions and Future Work

This chapter recalls the context of the thesis, provides a summary of the main contributions of this dissertation and also outlines the future research directions in the area of person re-identification. In this thesis, we have proposed effective solutions to the problem of person re-identification considering both occluded and unoccluded situations. Although several techniques have been developed in the past to address the person re-identification task, these suffer from limitations such as high response time, inability to deal with situations where subjects have similar appearances due to wearing similar-colored clothes, etc. We improve the state-of-the-art and develop effective methods for person re-identification that can be conveniently deployed at surveillance sites to track individuals seamlessly as they walk through a zone monitored by multiple cameras. Since it is relatively easy to control the flow of people at the entry/exit points of surveillance zones such as movie halls, stadiums, airports, and railway stations, we propose to carry out human tracking through re-identification by positioning the surveillance cameras above the entry and exit points of these surveillance zones to capture the frontal or near-frontal view of subjects as they walk in or out of the zone. The information gathered by the camera positioned at the entry gate/s of a surveillance zone would form the gallery set for re-identification, and the re-identification would be carried out as a subject approaches the exit gate and leaves the zone.

Contributions

The important contributions through this thesis are summarized next.

In our first two contribution chapters, namely Chapters 3 and 4, we aim to come up with effective approaches to image and video-based re-identification that improve upon the state=of-the art. Specifically, in Chapter 3, spatial information-based feature extraction techniques through two different Siamese network architectures have been discussed. In the first approach, we extract multi-scale features by incorporating dilation in the convolutional layers of the Siamese network (namely, *SMSNet*). The use of dilation helps the convolution operation to capture a wide viewing range of the image which also helps in performing accurate re-identification but makes the network more complex due to employing a high number of trainable parameters. As an improvement, we next propose a hierarchical re-identification scheme that performs color-based matching as an initial step and next carry out re-identification using a less complex Siamese network (namely, *Siamese Complex Box*) without any dilation in its layers. The use of hierarchical classification helps in eliminating vastly dissimilar samples at the first level of the hierarchy so that the second stage of *SCB*-based re-identification can be carried out accurately on the reduced gallery set. The average re-identification accuracy given by this SCB-based hierarchical approach has been seen to be better than the *SMSNet*-based approach on different data sets (namely, *CUHK_01*, *CUHK_03*, *Market-1501*, and *VIPeR*) as can be seen from Table 3.7 of Chapter 3. Further, the benefits of employing the hierarchical classification approach follows directly from the results given in Table 3.6 where it has been shown that the use of color-based feature matching stage in the first stage of the hierarchy results in a significantly high re-identification accuracy of 85% on the IIT (BHU) Re-identification data set, which is better than the second best approach by [131] by about 3%. The robustness of the Siamese classifier used in this work is also verified from the grouped bar plot in Fig. 3.16, which shows only minor variation in the re-identification accuracy on the same test sets extracted from the *CUHK_01*, *CUHK_03*, and *Market-1501* data sets, after training the model five different times from scratch.

Next, we study the use of spatio-temporal features for video-based person re-identification in Chapter 4. There exist only a few related approaches in the literature that exploit this spatio-temporal information from sequential frames to perform re-identification using some Deep recurrent neural network (RNN) models. It may be noted that training such models require extensive data sets with sequence information which is not practically available always. Obtaining adequate ground truth information about each subject for training a Deep RNN-based model is an extremely challenging task in surveillance scenarios where mostly videos of persons captured by one camera would serve as the training set. It appears that rather than relying on a single RNN model, an ensemble of different RNN-based models can lead to a better performance. Hence, in our next work we propose to use an ensemble model which we term as the Temporal Motion Aware Network *(T-MAN)*. The *T-MAN* is an ensemble of (i) a Full-Body Pose Attention Network *(FPAN)*, (ii) a Motion Pooling Network *(MPN)*, and (iii) a Long-short term Memory *(LSTM)*. The models are trained separately using the ground truth and during deployment the predictions from the three networks are combined to make a final conclusion about the class of a subject.Experimental results presented in Tables 4.1, 4.2, and 4.3 show that the proposed ensemble model performs with a higher re-identification accuracy (or, map) than most of the other existing approaches. This is since the fusion of the three above-mentioned models in *T-MAN* helps in capturing the different spatio-temporal aspects of motion from the same input video sequence, thereby enabling the extraction of more discriminative features. In this chapter, we have also experimentally verified the fact that the employment of ensemble of spatio-temporal features (as in *T-MAN*) is more effective for carrying out person re-identification from sequential frames instead of using spatial features only. The results in Table 4.6 also show that spatio-temporal feature-based re-identification through *T-MAN* is more effective in handling scenarios where subjects have similar-clothing conditions as compared to the spatial feature-based re-identification approach proposed in the previous chapter, i.e., Chapter 3 . The robustness of the proposed *T-MAN* architecture against different initialization of model parameters has also been tested through the grouped bar plot shown in Fig. 4.4. It has been observed that the model provides quite stable performance in all the different runs with different network initialization parameters. Based on the extensive set of experiments, we conclude that the hierarchical approach with the *SCB*-based classification stage (as proposed in Chapter 3) and the ensemble of RNN-based models (as proposed in Chapter 4) are the best approaches to perform re-identification from non-sequential data and sequential data, respectively. In the subsequent chapters, namely Chapters 5 and 6, we extend our work in Chapters 3 and 4 by handling the challenging occlusion situation in surveillance images/videos.

First, in Chapter 5, we target to solve the problem of person re-identification in the presence of occlusion given a set of non-sequential frames as input. Although there exist a few techniques in the literature that have provided solutions to this problem, most of these make use of a single end-to-end Deep Neural Network to predict the class of a subject directly from the input occluded frames. It appears that employing two separate dedicated Deep Networks for occlusion reconstruction and person re-identification can improve the overall re-identification accuracy further which we have studied in this chapter. Specifically, we examine the applicability of two different Deep Learning models, the first of which is based on a U-Net architecture with skip connections in between the convolution and the deconvolution layers and is termed as *OHGAN*, and the second is based on standard *Convolutional Autoencoder* that is devoid of any skip connections. Both the models are trained on a data set formed by synthetically occluded frames and their corresponding ground truth unoccluded frames, and their performances are evaluated on unseen synthetically occluded frames. A comparative study of the two reconstruction models by means of Dice similarity score in Table 5.3 indicates that the *Autoencoder*-based reconstruction generates better quality frames than the *OHGAN*-based reconstruction. This is since the skip connections present in the *OHGAN* tend to retain the impressions of the occluded patches present in the input image. This problem is likely to intensify if the input images are heavily occluded. The *Autoencoder*-generated images are next fine-tuned using a *DCGAN* and next as the a classification model we study the performances of both the *PCB* and the *SCB*, as discussed in Chapter 3. We have made a thorough comparative study of several occlusion handling methods using the *CUHK_01*, *CUHK_03*, and *Market-1501* data sets in Tables 5.6, 5.7, and 5.8, respectively, and observe that among these different approaches, our proposed one using *Autoencoder+DCGAN* for occlusion reconstruction and *SCB* for re-identification provides the best re-identification accuracy. However, due to the use of two separate Deep Learning models, the response time of our approach is expected to be higher than the existing single model-based re-identification techniques.

Next in Chapter 6, we consider the problem of person re-identification in the presence of occluded video sequences with sequential frames. To the best of our knowledge, there does not exist any Deep Neural Network-based occlusion reconstruction technique in the literature of person re-identification from occluded videos. It appears that the spatio-temporal information from a few preceding consecutive video frames can be utilized suitably using RNN-based models to predict the information in the current frame. Such a method can reconstruct a frame even if it is entirely occluded but a few previous frames are available, which the *Autoencoder*-based occlusion reconstruction technique discussed in Chapter 5 would have failed to do. In this work, we employ a time-series model based on *Convolutional LSTM* has been used to reconstruct the missing or occluded frames in a video sequence. This model exploits the spatio-temporal information from a few previous frames to predict the occlusion-free version of the present occluded frame. The qualitative and quantitative reconstruction results provided by *Conv-LSTM* fine-tuned with *DCGAN* as shown in Fig. 6.3 and Table 6.2 are indeed appealing. The robustness of the proposed *Conv-LSTM*-based reconstruction model has also been verified from the Box and whiskers plot in Fig. 6.2. Next, a comparative study is made with the *SCB*-based and *T-MAN*-based techniques proposed in Chapters 3 and 4, and also with some existing techniques on synthetically occluded *IIT (BHU) Re-identification* data and *PRID2011* data in Table 6.3. It has been seen that in general, our *T-MAN*-based method performs better than these existing techniques. Finally, we attempt to provide an unified interpretation of results by considering the best approaches determined in Chapters 3 to 6 by conducting a few additional experiments. First, we observe the re-identification accuracy and the response times for each of the proposed methods on video-based data sets, namely *iLIDS-VID*, *PRID2011*, and *IIT(BHU) Re-identification* data for varying degrees of occlusion in Table 6.5. We also conduct a similar experiment by considering three image-based data sets, namely *CUHK_01*, *CUHK_03*, and *Market-1501* corrupted with varying degrees of occlusion and present in the results in Table 6.6. In general, the approach proposed in Chapter 5 has been seen to perform well on image-based data sets but not so well on video data sets, whereas that proposed in Chapter 6 provides a more or less consistent performance for all types of data sets. This consistent re-identification performance may be due to the employment of the *T-MAN* ensemble during the classification stage, since ensemble classifiers are known to be more robust than single model-based classifier. As expected, the accuracy of the methods proposed in Chapters 3 and 4 are not appreciably high on occluded data sets since these do not contain any occlusion reconstruction module.

From the extensive study in the thesis, it can be concluded that the thesis indeed extends the state-of-the-art research on Computer Vision-based person re-identification. Improved solutions to both image and video-based person re-identification have been proposed and these are next fused with novel Deep Learning-based occlusion reconstruction models to carry out person re-identification in the presence of occlusion effectively. The methods proposed in Chapters 3 and 4 can be potentially applied in constrained surveillance setups without occlusion, whereas those proposed in Chapters 5 and 6 are suitable for more practical scenarios, where the input frames may be corrupted with occlusion. Our thesis also provides new and effective Deep Learning-based solutions to both image and video-based occlusion reconstruction which can have immense application in several other domains of Computer Vision such as gait recognition, human tracking, etc. The trained models have also been made publicly available for further comparative studies.

Limitations and Future Directions

The thesis deals with re-identification from two different types of inputs: while the approaches in Chapters 3 and 5 focus on developing effective methods to handle non-sequential image frames, that in Chapters 4 and 6 focus on sequential image frame-based re-identification. We have also experimentally verified that the use of spatio-temporal features is more effective to perform frame reconstruction or re-identification if the input frames are sequential in nature. However, in surveillance setups, typically large volumes of sequential data get captured by a network of cameras, in which several frames may contain irrelevant or useless information for re-identification. Also, storing and analyzing this video data require large volumes of storage which is quite cost-intensive. In such situations, carrying out re-identification using a set of randomly selected frames from the entire video sequence makes more sense than loading and working with the complete video data. The results presented in Tables 6.5 and 6.6 in the Chapter 6 highlight the fact that that the image-based reconstruction and re-identification method proposed in Chapter 5 suffers from the lack of consistency in prediction. While this approach performs with over 90% accuracy for each of the *CUHK_01*, *CUHK_03* and *Market-1501* data sets even for 30% occlusion (refer to Table 6.6), its performance on *iLIDS-VID*, *PRID2011*, and *IIT (BHU) Re-identification data* is not appreciably high even for 20% occlusion. In contrast, the video-based reconstruction and re-identification approach proposed in Chapter 6 has been seen to provide a more stable and accurate performance for each of the different data sets mentioned above. In future, more attention needs to be given to improve the prediction consistency of these image-based reconstruction and re-identification methods. A possible directions that can be explored in this domain are: (i) employing pose-based analysis for capturing better discriminative features, and (ii) carrying out multi-modal feature fusion from face, color-appearance, and silhouette shape information.

The use of Deep Learning models has boosted up automation in almost all areas of Computer Vision including person re-identification. However, the complexity of the Deep network increases as we increase the number of layers, resulting in high storage requirement and also large response time which may limit the practical applicability of a re-identification algorithm. As can be seen from the final column of the Tables 6.5 and 6.6, the overall time to re-identify a subject (i.e., the total time required to reconstruct the frames and re-identify by comparing with the gallery subjects) varies in the range 74-118 milliseconds for the method *M3* (i.e., the approach proposed in Chapter 5) and 70-105 milliseconds for the method *M4* (i.e., the approach proposed in Chapter 6). The response times are a bit high considering the fact that the gallery sets for each of the data sets considered in the thesis (except *Market-1501*) is not sufficiently high. For practical deployment, the reconstruction and re-identification models proposed in the thesis must be optimized by employing suitable network Compression strategies to come up with a lightweight network having less trainable parameters.

The hierarchical scheme for person re-identification discussed in Chapter 3 performs color-based clustering of gallery subjects as an initial step to reduce the search space for the Deep Siamese feature-based matching. The K-Means clustering method adopted in this work in the first step of the hierarchy can be replaced with a suitable adaptive clustering scheme which is likely to improve the overall rank-based re-identification accuracy further. The neural ensemble scheme described in Chapter 4 works by fusing the predictions from different sub-networks. This method can be extended by incorporating saliency detection methods to extract the important frame-level information from an input sequence before computing the features for re-identification.

Throughout the thesis, we propose solutions to several challenges in person re-identification by considering closed-set scenarios where the gallery is kept fixed. Our approaches need to be suitably extended to perform open-set re-identification with evolving gallery set, which is an important scope of work in this domain.

Further, as mentioned in Chapter 6, there do not exist any extensive video data sets with real occlusion in its sequences. Constructing one such data set and evaluating our proposed occlusion reconstruction models described in Chapters 5 and 6 must be done to observe their effectiveness in handling real-life occlusion scenarios.

Carrying out multi-sensor-based person re-identification is also another important direction of work. There can be situations where the gallery images are captured in daytime using RGB cameras, whereas the probe images are captured at night using thermal cameras. In such scenarios, color-based feature matching cannot be applied for effective re-identification. Rather, useful shape-related information needs to be extracted and matched, which is also another scope of future work.