# Abstract

Person re-identification refers to the process of finding one-one correspondences among images/videos of individuals captured by different cameras which may have overlapping/ non-overlapping fields of view. It plays a central role in tracking and monitoring crowd movement in public places, and hence it serves as an essential means for providing public security in surveillance sites. In this thesis, we target to come up with plausible approaches to Computer Vision-based person re-identification that can be conveniently deployed in surveillance setups where the movement of multiple persons is monitored by a network of cameras. In Chapter 1 of the thesis, we introduce the problem of re-identification, the challenges involved, along with the motivation of the present work, and main contributions of the thesis with highlights. Next, in Chapter 2, we present a thorough literature survey on person re-identification starting from the traditional contextual and non-contextual approaches to the modern Deep Learning-based approaches. In this chapter, we also present a thorough insight of the trend of research in the domain person re-identification by highlighting the summary and limitations of the recently published work in tabular form, from where we figure out the scopes for further research in this area.

It has been observed from the literature survey that the initial approaches to image-based person re-identification mostly consider color-based appearance descriptors for matching, whereas the modern approaches employ deep features to make the prediction more accurate and robust. While the initial approaches are passive and not so

reliable in the presence of varying lighting conditions or varying scales of the captured images, the modern Deep Learning approaches suffer from the use of large number of parameters that makes the process time-intensive specially if the gallery set is very large. The use of multi-scale features for person re-identification or fusion of the passive methods with Deep Neural Network-based methods is expected to significantly improve the overall effectiveness of re-identification, which we have studied in Chapter 3. Here, we propose two approaches to image-based person re-identification that deal with the extraction of effective spatial features from images through the use of (i) a multi-scale feature generation technique, and (ii) a hierarchical combination of color-based and Deep Siamese network-based features. We make a thorough comparative study among these two proposed techniques and also other state-of-the-art techniques and observe that both these outperform the existing approaches in terms of accuracy. Also, among the two, the second approach has been seen to provide a more consistent performance across different data sets and is less time-intensive due to following a hierarchical classification scheme.

Although there exist several re-identification techniques that work with videos/set of sequential frames, these all depend on a single model prediction. However, since video data sets are less extensive compared to image data sets, prediction from a single model may not be reliable. Hence, we propose to employ an ensemble of recurrent network models for the prediction so that the different spatio-temporal aspects of the motion data can be exploited for re-identification. Our proposed ensemble architecture is discussed in depth in Chapter 4, which combines the predictions from a Full-Body Pose Attention Network, a Motion Pooling Network, and a Long-Short Term Memory Network to re-identify an individual for a set of gallery subjects. Through extensive experiments and comparative study, we observe that fusion of the spatio-temporal information extracted by these three sub-networks helps in performing accurate re-identification from video data. We also observe that the use of spatial features alone

is also not so effective in situations where subjects are engaged in some sequential activities like walking, running, etc., and also in situations where subjects have almost similar clothing conditions.

The images/videos captured by the cameras in a surveillance zone are usually corrupted with occlusion caused by other static/dynamic objects present in the scene. To the best of our knowledge, although there exist a few Deep Learning-based occlusion reconstruction strategies in the context of person re-identification, none of these consider occlusion reconstruction and re-identification as two separate modules. Rather, these methods train a single Deep Neural Network to perform re-identification directly from the input occluded frames. It appears that the effectiveness of these approaches can be improved by training two separate dedicated Deep Neural Network architectures for occlusion reconstruction and re-identification and stacking them during deployment as a single end-to-end model. In Chapter 5, we have proposed two such improved techniques that reconstruct the occluded frames by employing Deep Neural Network generators, one of which is based on *UNet+DCGAN* with skip connections between the convolution and the deconvolution layers, while the other is based on *Autoencoder+DCGAN* without any skip connections. Following the reconstruction phase, another Deep Learning classifier is used for re-identification. We make a rigorous comparative study between the two proposed techniques and observe that the network formed by stacking *Autoencoder+DCGAN* performs the best between the two. Classification of the reconstructed images using a Siamese Network-based classifier shows that our proposed method outperforms the existing person re-identification techniques working with occluded sequences.

It may also be noted that, none of the existing techniques that handle occlusion in the person re-identification task are capable of exploiting the available spatio-temporal information if the input is a video sequence rather than a set of non-sequential frames. Due to relying on spatial pixel-based information only, the reconstruction quality of

these existing methods is poor in case a frame of the input sequence is heavily occluded. This limitation of existing techniques can be overcome by effectively utilizing the spatio-temporal information present in the adjacent sequential frames of a video sequence while making prediction about the missing/occluded frames, which we have considered in Chapter 6. Specifically, we propose an algorithm for occlusion reconstruction from videos by employing a *Conv-LSTM*-based generator and a *DCGAN*-based fine-tuner. The reconstruction and re-identification results given by our method on video data sets corrupted with occlusion are quite good and also outperform the related approaches for most experimental settings.

It may be noted that apart from *PRID2011* and *iLIDS-VID* data, most existing re-identification data sets do not consist of frames with a sequence of activities. To test the effectiveness of diverse video-based re-identification data sets, we construct another indoor data in our laboratory with frontal walking videos from 41 subjects and use it to evaluate the performances of the different approaches proposed in the Chapters 3-5 as well as for carrying out detailed comparative studies. This data set has been termed as the *IIT (BHU) Re-identification* data set and it has been made publicly available to the research community for further comparison. In Chapter 6, we also conduct an experiment to present an unified interpretation of results of all the approaches discussed in Chapters 3-6 using both the image-based and video-based occluded re-identification data sets. Our observation is that the approach proposed in Chapter 5 is most suited to carry out re-identification from non-sequential frames, while that proposed in Chapter 6 is suited for dealing with sequential frames captured by surveillance cameras in most real-life surveillance sites. In a more constrained setup, where clean input images/videos of a target subject are available, the approaches discussed in Chapters 3 and 4 can be conveniently used. Each of our trained models has been made publicly available to the research community for further comparative studies. Finally, in Chapter 7, we conclude the thesis and give insights to some future directions of work in the area of

person re-identification.

**Keywords:** *Image and Video-based Person Re-identification, Siamese Convolution Box, Temporal Motion Aware Network, Generative Modeling, Occlusion Reconstruction, Autoencoder, Convolutional LSTM*