Person Re-identification using Deep Learning

Smita Mishra (Post Graduate student)[1], Prof. Dhaval A. Parikh[2]

[1,2]*Department of Computer Engineering, Government Engineering College, Gandhinagar, 382028, Gujarat, India*

***Abstract*— In recent years, person re-identification has received much importance in computer vision field and became one of the most interesting subjects. Given an input image of a person captured from one camera, the task is to find out all the similar images of same person from the dataset. It is a difficult, because of the many issues such as different background appearances, occlusion problem and changes in view point. However, with the help of deep learning technology such as Convolutional Neural Network (CNN) extracting local features from images are extracted and analysed. In this paper we have proposed Deep compositional model that extract the features from person image and classify it to nearest similar images. We have utilized pre-trained fine tuneResNet-50 model trained on different ImageNet objects to classify the person re-id dataset. Ranking accuracy measures have been used on dataset to compute the similarity. Our trained fine-tune Resnet50 model can be used to perform person search or clustering as well. Here we have experimented on CUHK03 dataset and with task specific training our model achieves 93.75% Rank-1 accuracy, 94.01% Rank-5 accuracy and 94.43% Rank-10 accuracy.**

***Keywords*— Computer vision, convolutional neural networks, person re-identification, person matching, resnet50**

I. INTRODUCTION

Person Re-identification is the process of finding a similar person images are from large gallery set. Given a query image of a person the task is to find all similar images of the same person from the gallery of images if available. Person re-identification task performs the matching of same person images over multiple and non-overlapping camera views.

Figure 1 Person re-identification input and output

Person re-identification can apply in many security applications of video-surveillance, e.g., on-line tracking of person and retrieval of the video frames containing the query image person. It is applicable for surveillance system in public places where we require to monitor various locations and the behaviour of people in those areas. Events such as terrorist attacks and riots have occurred more frequently in recent years, this it is required for video network systems to improve the safety of people. Moreover, in many public transport places such as airports, stations, markets etc., video surveillance system has proven to be a useful application for preventing violent situations. Person re-identification process plays a vital role in processes that requires activity analysis and event recognition. In an intelligent video surveillance system, sequence of real-time video frames is grabbed from their source, normally closed-circuit television (CCTV) used and processed to extract the relevant information.

Problems related to person re-identification make it more difficult than the normal identification task as here we try to match the two similar person images which may have taken from different background. Some of the major problems are as following:

1. Detection issue: Prior to person re-identification the system has got to detect the person and define the bounding box of the person in a picture. The human body is very deformable in structure so detecting such deformable objects is a challenge in itself.
2. Illumination changes: Different daylight intensity, shades, reflection from different surfaces and other lighting can results in a different subject to seem in several shades and colours across cameras.
3. Low resolution: Several aged CCTV systems are with cameras of low resolution. Due to the shortage of data person Re-ID becomes even harder.
4. Occlusion: In crowded environments partial or maybe complete occlusion of persons by others presents challenge in extracting features.

It is required to develop techniques which will process these frames to extract the specified data in an automatic and operator-independent manner, this process is crucial for applications of surveillance systems.
Several Convolution Neural Network (CNN) are used to provide the solution for this task. In this paper, we have used the pre-trained model RESNET50 with some fine-tuning parameters to perform the person re-identification task. The model has trained on image dataset CUHK03 and tested for testing samples.

II. Methods and Materials

In the existing work many approaches have taken to extract features from the image based on this papers, Person re-identification models can be categorized in following ways:

A. Identification model

The identification deep models regard the person re-identification task as a classification issue, which outputs the corresponding labels of the input person images' attributes. The essential deep architecture of the identification model enriches the convolutional neural network features, Wu et al. [1] propose a fusion feature network (FFN), which mixes a spread of hand-crafted features (e.g., colour histogram features and texture features) and CNN features. within the backpropagation phase, the CNN features are constrained by the variability hand-crafted features. the general network is trained by SoftMax loss. For the reduction of variations while increasing the differences of inter-personal, in [2], a hybrid deep network is proposed for person Reid during which the low-level descriptors including colour histograms and SIFT are integrated into the long vectors, then that vectors and a deep neural network are combined to supply the finally non-linear features to represent person images

B. Verification model

This model takes a pair of images as input and display the similarity value [3] to find out whether the paired images are belonging to same person or not. Generally, verification model treats person Reid as a binary-class classification problem [3]. Li et al. [4] first introduce the verification model to deal with the person Reid problem. They propose a filter pairing neural network (FPNN) which incorporates max-out pooling layers and patch-matching to jointly handle geometric transforms and photometric, misalignment, background clutter and occlusions. within the same year, Yi et al. [5] present a "Siamese" deep network for learning metric evaluation. The network includes three shared parameters which are independent of each other in convolutional networks that perform on three non-overlapping parts of the 2 images. The deep descriptors of the 2 input images are produced by the fully connected layers, then the space of two output descriptors is calculated by the cosine function.

C. Distance metric-based deep model

It is more powerful at low lexical overlap. This model objective is to make small distance between the same person images and making large distances between different person images [3]. Triplet-based deep architecture is the most frequently used metric approach. Triplet model is first proposed by [6] to target the image retrieval task, and then Schroff et al. [7] introduce this triplet model into the face recognition community. Ding et al. [8] first adopt triplet model to address the person Reid task. Overall, distance metric-based deep model can dig the correlation between different person images and learn a similarity measurement within the training phase. So, the training target is consistent with its test manner. But the model needs to construct triple or quadruple units as input and repeatedly computes all possible triplets to select all the useful units for network, which reduces the training efficiency. In addition to this, the model uses weak label data of Reid datasets [3].

D. Part-based deep model

There are some literatures such as [3] adopted in part-based strategy to collect independent features for Reid. Cheng et al. [36] used the convolutional features for learning part-based features. After merging the worldwide and part-based feature vectors to provide the final features vector which is almost similar like [8], Li et al. [9] also division strategy. They used the same multi-classification losses function to optimize the model. In paper [3], the input image of person was first reshaped to 128x64 pixels and then divided into three parts, each of size 64x64 pixels. After then they used these three different parts to learn their individual features and used a fully connected layer to conclude these three different features into a single vector.

III. Proposed Methodology

In this paper we have trained a pretrained model ResNet50 such that they produce comparable feature vector. A pre-trained network has already learned to extract powerful and informative features from the natural images and therefore the weights already fixed for the actual application. To perform this task our model is trained by deep layer model on CUHK3 datasets.

E. Gradient Vanishing Problem:

In shallow network, where only a few layers are used, in that case activations were not a big problem. However, in deep neural network where more layers are used, it causes the problem of gradient to be too small for training the network to work effectively.

F. Residual Network:

In order to solve the problem of the vanishing gradient in deep neural network, a new architecture introduced the concept called Residual Network. In this type of network architecture, we use a technique called skip connections. The skip connection bypass the training from a few layers and connects input directly to the output. While in traditional neural networks they try to learn a mapping function, but a residual layer attempts to approximate output function by adjusting identity function. In this way, it tries to makes output function zero so that input becomes the output.

G. Fine-tune Resnet50

In the original residual module structure, the model accepts an RELU activation map as an input and applies a series of operation prior to adding output to the original input. In this method, RELUs layer and batch normalizations layer are placed before the convolution layer and hence this module called as pre-activate residual model. This is the different approach of layering compare to normal one where RELUs and batch normalizations layers are coming after the convolution layer. The basic flow diagram of proposed work is as shown below:

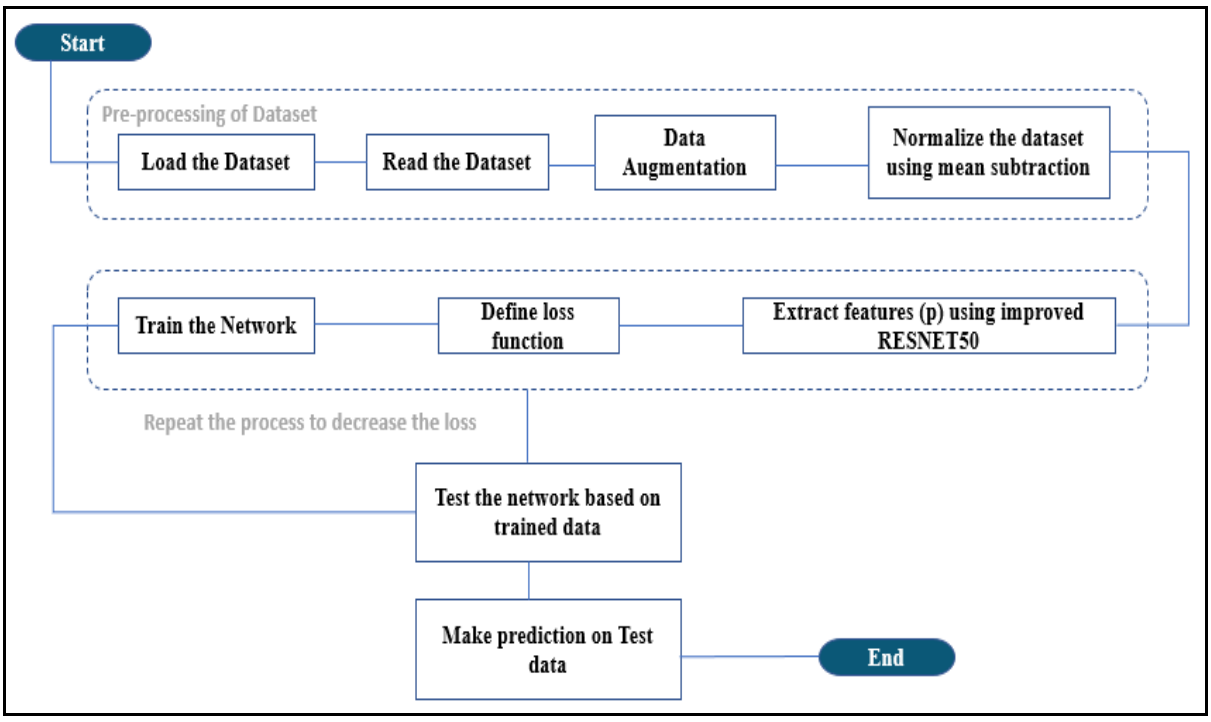

Figure 2 flow diagram of proposed work

The entire workflow starts with pre-processing of dataset. In pre-processing of dataset mean subtraction perform as like in ImageNet dataset, so that model can be trained smoothly. Data augmentation is also used to create a greater number of images for single images using keras libraries to increase the strength of dataset.

Following is the algorithm for above mention process:

Step 1. Divides the original dataset into two set training and testing sets in 70% and 30% ratio, respectively.
Step 2. Arrange the images of person in directory wise folder i.e, placed all images of one person into one folder, to use Keras' ImageDataGenerator class and flow_from_directory functions.
Step 3. Perform preprocessing od data using mean subtraction and data augmentation process to increase number of images.
Step 4. Then load the ResNet50 model using the pre-trained ImageNet weights without the fully-connected layers as this needs to change according to person Reid task.
Step 5. After then, applied an average pooling layer with 7x7 convolution and then appended fully-connected layers
Step 6. Trained the network for specific number of epochs. Evaluate the test samples for finding accuracy and loss.

For optimization of model Adam algorithm used for stochastic gradient descent for training deep learning models. It combines the properties of the AdaGrad and RMSProp algorithm and provides an optimization algorithm that is able to handle sparse gradients on noisy problems. It is relatively easy to configure and the default configuration parameters do well on most of the problems.

H. Evaluating Performance

Most methods have used the Rank-n accuracy and mAP metrics to evaluate the result. We also used the same metrics in our experiment to evaluate the output.

1. Rank-n measures:

When a query image is given, the model must retrieve the predicted matching image from the data. If the first image in output is the one, we searching then it namely as rank-1. It is difficult to find specific similar samples from a large amount of gallery, and few algorithms can achieve the highest precision. Therefore, as long as the correct picture is retrieved from the top n result, the accuracy can be calculated as rank-n. Generally, for evaluation the often-used rankings are rank-1, rank-5 and rank-10.

2. mAP (Mean Average Precision):

The mAP takes the person re-identification as a retrieval task and every query image has a matching cross-camera ground truth. For each query, the area under the precision-recall (PR) curve, which is the average precision is calculated first and then take the mean of all the precision. There are many ways to calculate the area under the PR curve, one of which is shown as follows:

$$AP = \frac{1}{2} \sum_{i=2}^{M} \left((recall_i - recall_{i-1}) \left(precise_i + precise_{i-1} \right) \right)$$

IV. RESULTS

Dataset here used is CUHK03 which is collected from The Chinese University of Hong Kong (CUHK) college of China. This dataset consists of 14,097 images of 1,467 different identities or person images, where 6 campus cameras were deployed for image collection and each identity is captured by 2 campus cameras. We have initially trained the model for 5 epochs, the output we got was 40%. Gradually we increase the number of epochs by 5. The model is trained for 35 epoch and at epoch 33 we getting highest accuracy of 95.54%. The following graph shows the training and validation accuracy and losses when model trained for 35 epochs. It seems that till epoch 33 the accuracy is 95.54% then after it remain stable. The graph shows that accuracy increases and loss getting decrease in validation phase. Also, as number of epochs increases the loss getting degraded.

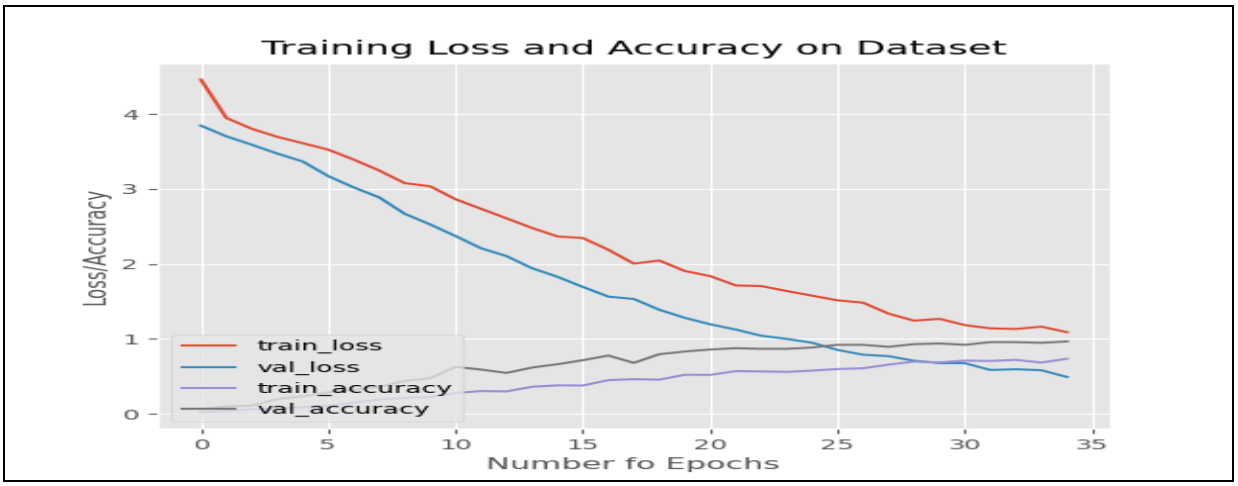

Figure 3 Training accuracy and losses graph for 35 epochs

Once the model is trained it saved as PersonReidmodel.h5 file so that same can be used for testing. Following figure shows the testing accuracy of model tested on few samples.

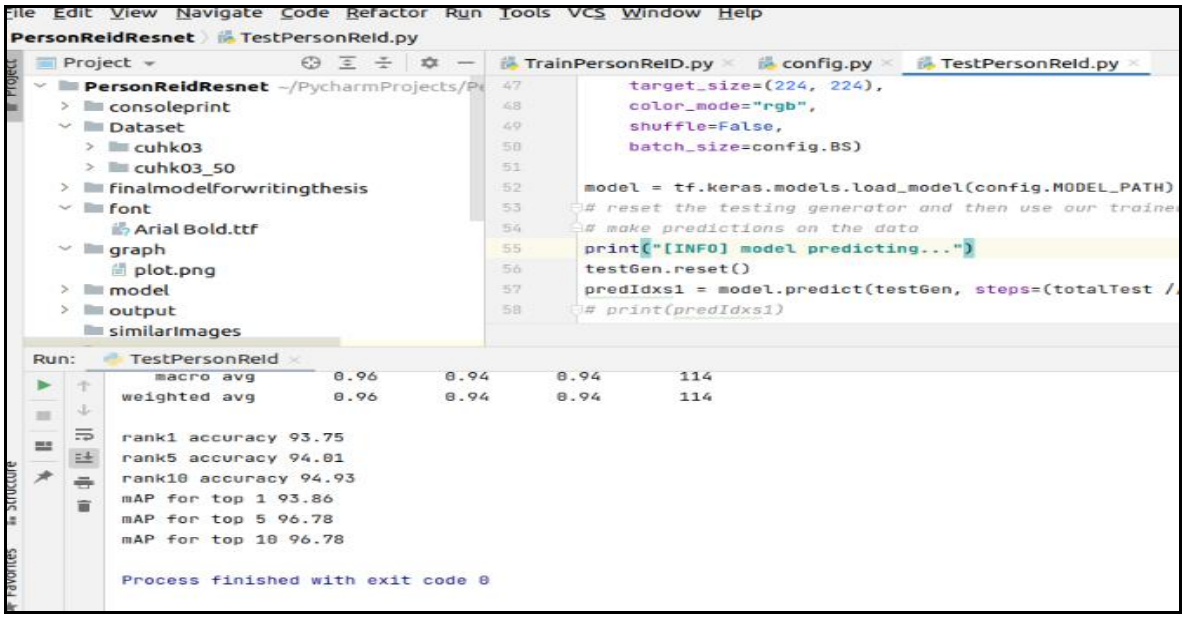

Figure 4 Saved model tested on testing sample

As shown in figure 3, the accuracy gets stable at epoch 33 after then it starts reducing after epoch 35. So, the highest accuracy we have achieved is 95.54%. We have also considered the rank-5 and rank-10 accuracy for our model and we got the following observation:

TABLE 1 RANK AND MAP VALUES OF PROPOSED MODEL

Parameters	**Observation on CUHK03 dataset in (%)**
Rank-1 accuracy	**93.75**
mAP for k=1	93.86
Rank-5 accuracy	94.01
mAP for k=5	96.78
Rank-10 accuracy	94.93
mAP for k=10	96.78

The following table shows the rank accuracy of Person Reid fine-tune ResNet50 model on CUHK03 dataset comparing with existing approach as:

TABLE 2 ACCURACY COMPARISON OF THE PROPOSED METHOD

Model	**rank-1 (%)**	**rank-5 (%)**	**rank-10 (%)**
SSM *[15]*	76.6	94.6	98
Spindle *[14]*	88.5	97.8	98.6
DPAR *[13]*	85.4	97.6	99.4
DLMR by Chen et. al. *[12]*	86.7	-	-
HydraPlus *[11]*	91.8	98.4	99.1
Inception-V3^{ft} *[10]*	88.73	97.82	98.94
Inception-V3^{ft*} *[10]*	92.81	98.9	99.35
Fine-tune ResNet50 Person Reid	**93.75**	**94.01**	**94.93**

From the table it concluded that fine-tune ResNet50 model showing improvement on accuracy by ~1% in rank-1. We also tried to show rank visualization using PIL and imutlis libraries. Following figure show the visualization of output, the top ranked images predicted by trained model. The green marked shows that the model predicted the correct person while red shows the wrong prediction. The numpy libraries for sorting the top-k prediction of model and showing the randomly selected single image for displaying the output. The figure 5 shows the top-10 visualization of output images predicted by model for given query image.

Figure 5 Rank Visualization of person re-identification model

Similarly, to see all the similar images of the same person then we implemented the code which display all the images of person available in gallery from the actually class predicted by model correctly in top-10 predictions.

Figure 6 Similar images of query person predicted by model in top-10 list

This are the similar images of query person predicted by model in top 10 rank classes. The first one is the query image in black border while images in yellow border are the correct images of person predicted by our model.

V. CONCLUSIONS

In this work the model is trained matching the two person images for doing person re-identification. We have trained fine tune deep neural network ResNet50 by changing its head layer with the sole purpose of generating feature vector of person. The highest accuracy we got is rank1 accuracy 93.75%, rank5 accuracy 94.01% and rank 10 accuracy 94.93%. This model will be useful in surveillances system for tracking the person and strengthens the security. This paper also contain reidentification experiment for the proposed algorithm, compares it with the main algorithm, and obtains a good result. The experiment shows that the algorithm has a relatively improvement in the average accuracy and has an obvious advantage on rank-1.

REFERENCES

[1] S. Wu, Y. Chen, X. Li, A. Wu, J. You and W. Zheng, "An enhanced deep feature representation for person re-identification," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-8, doi: 10.1109/WACV.2016.7477681.

[2] Wu, Lin & Shen, Chunhua & Hengel, Anton. (2016). Deep Linear Discriminant Analysis on Fisher Networks: A Hybrid Architecture for Person Re-identification. Pattern Recognition. 65. 10.1016/j.patcog.2016.12.022.

[3] D. Wu et al., "Omnidirectional Feature Learning for Person Re-Identification," in IEEE Access, vol. 7, pp. 28402-28411, 2019, doi: 10.1109/ACCESS.2019.2901764.

[4] W. Li, R. Zhao, T. Xiao and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152-159, doi: 10.1109/CVPR.2014.27.

[5] D. Yi, Z. Lei, S. Liao and S. Z. Li, "Deep Metric Learning for Person Re-identification," 2014 22nd International Conference on Pattern Recognition, 2014, pp. 34-39, doi: 10.1109/ICPR.2014.16.

[6] J. Wang et al., "Learning Fine-Grained Image Similarity with Deep Ranking," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1386-1393, doi: 10.1109/CVPR.2014.180.

[7] *F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.*

[8] D. Cheng, Y. Gong, S. Zhou, J. Wang and N. Zheng, "Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1335-1344, doi: 10.1109/CVPR.2016.149.

[9] Li, Wei & Zhu, Xiatian & Gong, Shaogang. (2017). Person Re-Identification by Deep Joint Learning of Multi-Loss Classification. 2194-2200. 10.24963/ijcai.2017/305.

[10] Kalayeh, Mahdi & Basaran, Emrah & Gökmen, Muhittin & Kamasak, Mustafa & Shah, Mubarak. (2018). Human Semantic Parsing for Person Re-identification. 1062-1071. 10.1109/CVPR.2018.00117.

[11] Liu, Xihui & Zhao, Haiyu & Tian, Maoqing & Sheng, Lu & Shao, Jing & Yi, Shuai & Yan, Junjie & Wang, Xiaogang. (2017). HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. 350-359. 10.1109/ICCV.2017.46.

[12] Y. Chen, X. Zhu and S. Gong, "Person Re-identification by Deep Learning Multi-Scale Representations," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 2590-2600, doi: 10.1109/ICCVW.2017.304.

[13] Zhao, Liming & Li, Xi & Wang, Jingdong & Zhuang, Yueting. (2017). Deeply-Learned Part-Aligned Representations for Person Re-Identification.

[14] H. Zhao et al., "Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 907-915, doi: 10.1109/CVPR.2017.103.

[15] H. Liu, J. Feng, M. Qi, J. Jiang and S. Yan, "End-to-End Comparative Attention Networks for Person Re-Identification," in IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3492-3506, July 2017, doi: 10.1109/TIP.2017.2700762.