Person Re-Identification using Attribute Learning

Sai Eswar Epuri
saieswar@tamu.edu

Manoj Reddy Rupireddy
rupimanoj@tamu.edu

Janardhana Swamy Adapa
janardhan_adapa@tamu.edu

1. Problem Statement

Person re-identification is a fundamental task in automated video surveillance and has been an area of intense research in the past few years. The problem has received significant attention from the computer vision research community due to its wide applicability and utility in the field of video surveillance, robotics, multimedia and forensics. Re-identification (Re-ID) is defined as a process of establishing the correspondence between images of a person taken from different cameras. It is used to determine whether instances captured by different cameras belong to the same person, in other words, assign a stable ID to different instances of the person. The difficulties of solving this problem involve visual ambiguity and spatiotemporal uncertainty in a person's appearance across different cameras and often these difficulties are compounded by low-resolution images or poor quality video feeds with large amounts of unrelated information in them that does not help in re-identification. In this work, we aim to demonstrate that by combining the global level features of a target person with the low level semantics such as attributes can improve person Re-ID accuracy.

2. Literature Survey

For person re-identification problem, methods have been proposed that use unsupervised features matching techniques [2] where features are calculated using color histogram or SIFT descriptors. But with the availability of large labeled datasets such as Market1501 and DukeMTMC, CNN based person re-identification systems are widely getting used in the present research community. These deep network models broadly adopt two approaches, either by directly learning similarity metric using siamese networks[3] or using the classification networks to extract the feature embeddings[4]. Our work adopts the second approach, where a pre-trained deep neural network model of classification task was fine-tuned on a labeled dataset. The learned feature embeddings from the last layer are used to compute the similarity between the query and gallery images by using a distance metric. Though these deep networks extract the global features they did not account for varying lighting conditions, view angles of cameras etc. To address these challenges in image retrieval tasks, approaches have been proposed that employs techniques such as data augmentation, camera style adaptation and part pooling [7][8]. Similarly, to avoid retrieval errors in the cases where different individuals have almost same global descriptors such as height and attire, our work demonstrates how individual attributes information such as gender, hair length, lower body, and upper body clothing etc can be used as a auxiliary information in the feature embeddings learning process. In previous approaches, Person re-identification problem has been solved by using attributes alone [5] where the network is trained to extract multiple attributes of a target and similarity is calculated based on the distance between the detected attributes. This approach does not take advantage of the capability of deep networks to learn the global features embeddings and it completely relies on medium level semantics such as attributes. Our work demonstrates that by combining the loss functions of person re-id classification network[6] and attributes classification network during the training process, we can achieve better accuracy compared to re-ID models.

3. Technical Plan

In this project, we aim to perform Person Re-ID task by Attribute Learning using multi-task learning process. This multi-task method integrates an ID classification loss and a number of attribute classification losses, and back-propagates the weighted sum of the individual losses. We plan to use Market-1501 and DukeMTMC-reID datasets, as these datasets have attribute labels. In Market-1501, we have 27 labeled attributes, and in DukeMTMC-reID, we have 23 labeled attributes. For our NN architecture, we use ResNet-50 as the base network which was pre-trained on ImageNet. As mentioned in the figure-1, our architecture contains M+1 FC layers, in which M layers followed by softmax layers corresponds to attribute recognition, and one layer corresponds to identity classification. Here, M denotes the number of attributes for our dataset. FC layer corresponding to each attribute has its number of nodes equal to the number of classes of that attribute. This architecture also contains the dropout layer after the ResNet base network. We fine-tune this network using

the attributes and identity labels. In our proposed architecture, we have M+1 losses, in which M losses for attribute classification, and one loss for identity classification. The cross entropy loss for identity classification is

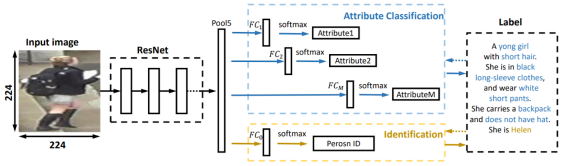

Figure 1. Overview of the architecture

$$L_{ID}(f,d) = -\sum_{k=1}^{K} log(p(k))q(k)$$

Here p(k) is the predicted probability of each ID label, and q(k)=1, where k is the ground-truth ID label. And the loss for each attribute prediction is

$$L_{att}(f,l) = -\sum_{j=1}^{m} log(p(j))q(j)$$

Here p(j) is the predicted probability of each class in the corresponding attribute, and q(t)=1, where t is the ground truth ID label. We use multi-task learning process to train this Neural Network. Given an input image, the network simultaneously predicts its identity and a set of attributes. And the final loss function is

$$L = \lambda L_{ID} + \frac{1}{M}\sum_{i=1}^{M} L_{att}$$

Here L_{ID} is identity classification loss and L_{att} is attribute classification loss. We aim to find the optimum value for which corresponds to best accuracy on the validation set. For testing purposes, we extract feature embeddings of gallery images and query images from the output of the last CNN layer and calculate the similarity between them by using L2 norm distance. Gallery images are ranked based on this distance metric and top-k ranked images can be retrieved accordingly.

The evaluation metrics are top-1 score - whether the top class with the highest probability is same as the target label, top-5 score - whether the target label is one of the top 5 predictions, and mAP (mean Average Precision) - average of maximum precision at all recall levels. We are going to use pytorch as a deep learning framework, and HPRC computing resources for running our models.

References

[1] Rui Zhao, Wanli Ouyang, Xiaogang Wang. Person Re-identification by Salience Matching

[2] Jie Guo, Yuele Zhang , Zheng Huang, and Weidong Qiu. Person Re-Identification by Weighted Integration of Sparse and Collaborative Representation

[3] Dong Yi, Zhen Lei, Shengcai Liao and Stan Z. Li. Deep Metric Learning for Person Re-Identification

[4] Tong Xiao, Hongsheng Li, Wanli Ouyang, Xiaogang Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification

[5] Ryan Layne, Timothy Hospedales, Shaogang Gong. Person Re-identification by Attributes

[6] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person Re-identification: Past, Present and Future

[7] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, Yi Yang. Camera Style Adaptation for Person Re-identification

[8] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, Yi Yang. Random Erasing Data Augmentation

[9] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu and Yi Yang. https://github.com/vana77/DukeMTMC-attribute

[10] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang, Improving Person Re-identification by Attribute and Identity Learning.

[11] Apurva Bedagkar-Gala, Shishir K. Shah. A survey of approaches and trends in person re-identification.

[12] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, Qi Tian. Scalable Person Re-identification: A Benchmark