Open-Set Metric Learning for Person Re-Identification in the Wild

by

Dibyadip Chatterjee
Roll: 001610701093

Arpan Bhowmik
Roll: 001610701079

Thesis submitted in partial fulfilment of the requirements for the degree of

Bachelor of Engineering

in

Electronics and Telecommunication Engineering

Jadavpur University
Kolkata - 700032

June 2020

Certificate

This is to certify that the bachelor's thesis entitled "**Open-Set Metric Learning for Person Re-Identification in the Wild**" submitted by **Dibyadip Chatterjee** and **Arpan Bhowmik** for the partial fulfillment of the degree of Bachelor of Electronics and Telecommunication Engineering of Jadavpur University is based on the assigned final year project work during the session 2019-2020 under the supervision of Prof. Ananda Shankar Chowdhury of Electronics and Telecommunication Engineering department of Jadavpur University, Kolkata, West Bengal, India.

Roll Number	Name of Student
001610701079	Arpan Bhowmik
001610701093	Dibyadip Chatterjee

Prof. Ananda Shankar Chowdhury
(Thesis Supervisor)

Date:

Abstract

Open-set learning is one of the most significant challenges of computer vision and no prior method exists that solves the problem of open-set in metric learning systems. In this work, we present a novel open-set metric learning (OSML) model to re-identity person in the wild. Person re-identification in the wild requires simultaneous detection and re-identification from non-overlapping raw video feeds. This is a more reasonable assumption in a real world where the gallery set has to be generated on a per frame basis. Person re-ID in the wild is essentially an open-set problem since the gallery set is dynamic (generated per frame) as a result of which a probe sample may not be present in the gallery at that moment. Close-set re-ID models are incapable of rejecting probe samples absent in the gallery and thus raises false alarms. Traditionally metric learning methods have outperformed classification models in re-ID, which led us to design our own open-set metric learning (OSML) model based on the concept of Large Margin Nearest Neighbor (LMNN) and weibull distribution. Our model, named Open-Set LMNN (OS-LMNN) has been tested exhaustively on the publicly available PRW dataset and has been compared with existing metric learning methods to demonstrate the robustness of our model.

Preface

A version of this work has appeared in the following publication:

- A. Sikdar, D. Chatterjee, A. Bhowmik, A.S. Chowdhury, "*Open-Set Metric Learning for Person Re-identification in the Wild*", 27th IEEE International Conference on Image Processing (ICIP), Abu Dhabi, UAE, 2020

Acknowledgements

We would like to extend our gratitude to a number of people who have been a continuous source of support during our bachelors and has directly or indirectly contributed to the completion of our thesis. First and foremost, we would like to thank our supervisor Prof. Ananda Shankar Chowdhury for his constant encouragement and support. From the very beginning of our research, he has been extremely supportive of our work and interests. Secondly, we would like to thank Arindam Sikdar for formulating the problem and designing hypotheses following which we arrived at our solution. This work would not have been possible without his substantial contribution. Next, we would like to thank the entire faculty of the Department of Electronics and Telecommunication Engineering, Jadavpur University, for providing us with the required knowledge and infrastructure throughout our four years in the department. Lastly, we would like to thank our families, friends and well-wishers whose support was crucial towards the completion of our bachelor's thesis.

Table of Contents

Chapter 1

Introduction

1.1 A brief background

Accurately identifying people from digital images and videos with the help of computer algorithms is a problem that has been extensively studied by researchers since 1960s. The earliest approaches by scientists on this front were to recognize human faces which involved the manual marking of various facial landmarks viz. eye-centers, mouth, nose, etc. and these were mathematically rotated by a computer to compensate for pose variation. The distances between the facial landmarks were also computed and compared between images to determine their identity.

Presently, facial recognition systems are used extensively in social media (e.g. Facebook, Instagram, etc.) for automatically identifying persons. Their algorithms learn the features of a person's face from manually tagged photos and look for the same features in new images.

A facial recognition system like Fig 1.1 is applicable to faces that are directed towards the camera with sufficient illumination such that the facial features are detectable. Now, for surveillance applications majority of the images/videos are available from CCTV footage where it is difficult to discern the facial features of a person. Furthermore, a person may deliberately hide

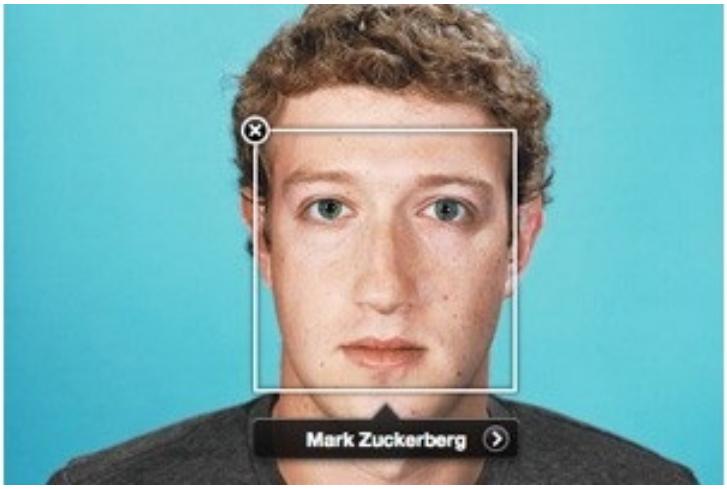


Figure 1.1: DeepFace — A deep facial recognition system created by researchers at Facebook. It identifies human faces from digital images.

his/her face from the CCTV camera. Thus given an image of a person we have to find a similar person from a gallery of images using not just the facial features but also the features extracted from the other parts of their body (e.g. height, colour of clothes, etc.)

1.2 Person re-identification and it's history

Person re-identification can be formally defined as: "Given an image or video of a person taken from one camera, re-identification is the process of identifying the same person from images or videos taken from a different camera with non-overlapping fields of views (FOVs)."

In other words we are assigning a stable ID to a person in a multi-camera setting. Initially the term person re-ID did not exist formally and was closely associated with multi-camera tracking. In such problems, the calibration of cameras having non-overlapping FOVs supplemented with an appearance model were used [2]. According to [3], the earliest work on multi-camera tracking with explicit person re-ID was proposed in [4] where a dynamic

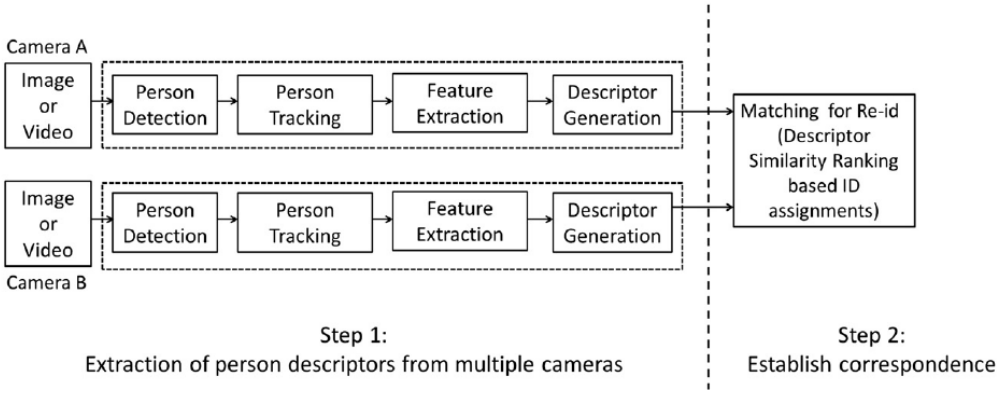


Figure 1.2: Person re-ID System Diagram [1].

Bayesian network was used to encode the probabilistic relationship between features and labels from tracklets. In 2006, the authors of [5] evaluated their work on a dataset with 44 persons captured by 3 cameras with moderately overlapping FOVs and it marked the independence of person re-ID as a separate computer vision problem.

Until 2010, all works on re-ID were based on matching images (one-shot). The first works on video based re-ID (multi-shot) were [6, 7] where frames were randomly selected from raw videos. They also showed that multiple frames per person improves re-ID performance significantly and that the performance saturates as the number of frames increases. All of these works deploy models based on color features and a standard distance metric. As deep learning became popular in computer vision, several re-ID models [8, 9, 10, 11] adopted them to increase their accuracy. Majority of the work use hand-cropped boxes or boxes produced by a fixed detector in their experiments but it is also important to study the impact of different pedestrian detectors on re-ID accuracy [12, 13] for performing end-to-end re-ID.

1.3 Challenges in person re-identification

There are several challenges related to person re-ID. The major ones arise due to the variation of the same person over different timestamps. For short-term re-ID i.e. re-ID occurring over hours, the changes in posture (standing, lying, sitting, etc.) across the different cameras become a major hurdle in re-identifying them. There might be variations in illumination across the different cameras i.e. one camera may be installed in a dark room whereas another one may be installed in a sufficiently lit place. Cameras might be installed in surveying over dense environments where a crowd of people are present and it is difficult to localize each person separately due to occlusion or due to strong similarity in appearance (posture, uniform clothing, etc.) For long-term re-ID i.e. re-ID occurring over weeks, the color of clothes and hair style may change in time. In such listed cases, the key features that we use to uniquely identify a person becomes the source of unreliability and hence careful measures are needed to be adopted to handle such issues.

1.4 Variants of person re-identification

There are many variants of person re-ID that arises as a result of variation in the definition of the problem. Some fundamentally important ones are mentioned as follows:

One-shot vs multi-shot re-ID:

A person re-ID system can have an image (one-shot) or a video (multi-shot) with a sufficient amount of frames as input. For one-shot re-ID, given an image, persons are localized within it and features are extracted for each of them which are then matched with another image to compute similarity measures among them. For multi-shot re-ID, given multiple images (frames) of a person as extracted from a video, each image can be treated separately and

one-shot re-ID may be performed. On the other hand, multiple instances of a person can be generated through *tracking* or the temporal dependencies between consecutive frames can be considered for better feature representation of the person.

Open-set vs close-set re-ID:

A re-ID is said to be open-set if the probe set may or may not be a subset of the gallery set i.e. the gallery might not contain the true ID of a probe sample. Practically, it means that in an open world scenario a probe camera might encounter a new person whose feature needs to extracted and added to the gallery. On the other hand, close-set re-ID is a constrained form of open-set where the probe set is always a subset of the open-set i.e. there are no probe sample IDs that are absent in the gallery. When it is guaranteed that the probe sample ID is present in the gallery we match it with the gallery samples and assign it the ID of the gallery sample it has the highest match with.

1.5 Person re-identification in the wild

In both cases of open and close-set re-ID, it is to be noted that the gallery has already been generated and fixed. But for person re-identification in the wild we are only provided with the raw video, thus detecting & localizing the persons and creating the gallery from the sampled video frames is also a part of the task. Hence person re-identification in the wild is an end-to-end system whose performance is affected by both the *pedestrian detector* and *person re-identification* methods. In such scenarios, all the probe persons may not be present in a particular frame and as a result the probe set is usually larger than the dynamically generated gallery set per frame. So a rejection mechanism is necessary to handle false alarms.

Chapter 2

Prior Work

The fundamental problem of person re-id is to compare a person of interest as seen by a *probe* camera to a *gallery* of candidates captured from another camera whose FOV does not overlap with the FOV of the probe camera. If a true match to the probe sample exists in the gallery, it should have a high matching score, or rank, compared to the incorrect candidates. As described in Sec. 1.4, re-ID can be broadly classified into image based (one-shot) and video based (multi-shot).

2.1 Image based person re-ID

Starting as a multi-camera tracking problem, person re-ID has been majorly explored using images in the past [5]. Several methods were proposed framed on a two-step approach — pedestrian description and distance metric learning. Commonly used pedestrian descriptors were based on color [5, 14, 7]. In [14], 8 color channels viz. RBG, HS & YCbCr were used along with 21 texture filters on the luminance channel followed by the partitioning of pedestrians into horizontal stripes. Such handcrafted features used for generating pedestrian descriptors have remained relatively same even in some of the later works [15, 16, 17] where color histograms were predominantly used.

Instead of using low-level features such as color and textures, some mid-level features based on attributes have also been proposed [18, 19].In [20] a large-scale dataset with richly annotated pedestrian attributes was collected to facilitate attribute based re-ID methods. In order to achieve a good performance with such handcrafted features, a good distance metric was also essential [21]. Popular metric learning models used in person re-ID include KISSME [22], LMNN [23], XQDA [24] and DNS [25].

With the advent of deep learning for computer vision [26], recent works tend more towards the usage of deep models due to transfer learning i.e. learning better representations from an image with a model pre-trained on a large dataset. Approaches are of various types but can be softly demarcated into classification based [26, 27, 28] and distance based siamese model using image pairs [29] or triplets [30] as input. In some works the deep features have been learnt end-to-end [9, 10, 31, 32]. Major issues in re-ID like scalability has been handled in [33] where a siamese network having convolution operations of different filter size was used. In other cases low-level descriptors like SIFT and color histograms have been augmented with fully connected layers with a redefined objective function that produces feature embedding having low intra-class variance but high inter-class variance [34, 35].

2.2 Video based person re-ID

Although person re-ID has been explored predominantly with images, in the recent years, video based multi-shot re-ID has been under spotlight and has become a popular topic of research. Videos can be used to garner multiple instances of a single probe sample that enables better feature representation. Earliest methods at attempting to solve re-ID from videos have adopted the use of handcrafted features based on foreground segmentation and color descriptors. [6, 7] have used methods related to one-shot techniques as described in Sec. 2.1, but the major difference lies in their multi-shot matching. In [36],

multiple shots are used to train a discriminative boosting model based on a set of co-variance features. Several works have been done on utilizing the underlying spatio-temporal structure of videos [37, 37]. [38] makes use of multiple instances for a person and propose that the feature representation of a probe sample can be a linear combination of samples with same ID in the gallery. Multiple instances of an ID can also be employed to enhance body part alignments like *pose*. Also, *Tracking* can be implemented to avoid occlusion in dense environments. In [39], the pose of a person is computed and frames having the same pose are matched with higher confidence. On top of that, temporal dependencies can also be incorporated to better represent a person over multiple shots [40, 41, 42]. [43] proposes simultaneous learning of intra and inter-video distances to create a more discriminative and compact video representation.

In case of image based re-ID a major bottleneck of using deep learning was the lack of available data i.e. one instance was generally available per probe ID. But in case of videos, multiple instances are readily available and hence deep learning approaches have flourished in such scenarios. Learning a single feature representation from a sequence of frames has been achieved through average/max pooling after passing the sampled frames through a few convolutional layers [44, 45, 46]. Aggregating frame features into a single compact vector yields competitive results but temporal information can also be harnessed [47, 48] to improve the performance even more. Computing poses and local deformations of a body part is learned in [49] where an automatic part alignment is done during the learning phase without extra supervision. Supervision is eliminated completely in [11] by incrementally learning the discriminative features required for re-ID from automatically generated person tracklets.

Chapter 3

Overview of Methodologies

3.1 Regions with CNN features (R-CNN)

Over the past few years, extensive work has been done on object detection from static images. Object detection differs from vanilla image classification in a sense that there can be multiple instances of a single object or multiple objects of different classes present within an image and the task is classifying all the objects present and drawing a bounding box describing the unique location of each object. One method is to slide a rectangular window with fixed parameters over the image and detect the presence of an object in every such window using a CNN. This method has an enormous computational overhead which is conciliated with the use of region proposals in R-CNNs [27].

R-CNN is essentially a two-stage object detector where the first stage is dedicated to compute region proposals from an image. Such category-independent region proposals are computed using selective search [50]. The second stage is a classifier that classifies each of the proposed regions into labelled objects of different classes (including background i.e. no object).

The original R-CNN was later augmented into faster and lighter models by the use of RoI pooling layer in Fast R-CNN [28] and region proposal network in Faster R-CNN [51]. Faster R-CNN was further improved to Mask

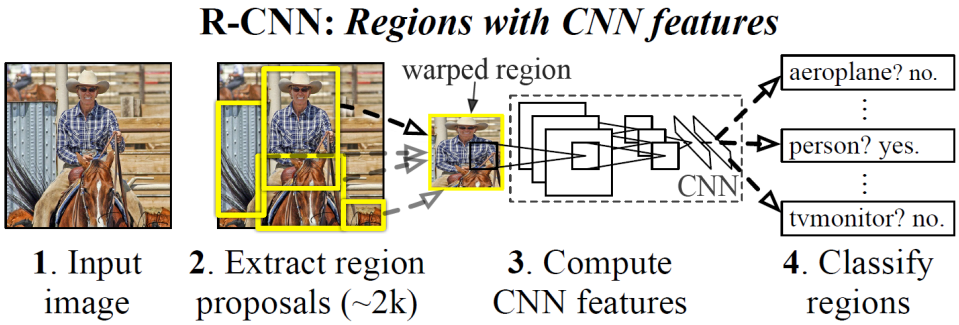


Figure 3.1: R-CNN pipeline

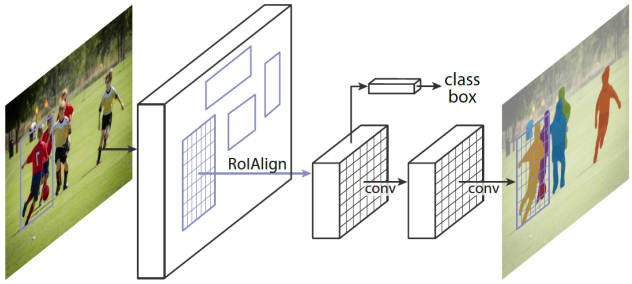


Figure 3.2: Instance Segmentation using Mask R-CNN with RoI Align

R-CNN [52] by introducing RoI Align. Mask R-CNN uses Feature Pyramid Network (FPN) [53] over it's ResNet [54] backbone that pools object RoIs combining feature maps of various CNN layers thus providing better results for detecting objects of varying scales especially small objects. On top of object detection, Mask R-CNN can also be used for instance segmentation.

3.2 Large Margin Nearest Neighbor (LMNN)

Calculating distances between data points is an important task in machine learning which is traditionally performed by a standard distance metric like

euclidean, manhattan or cosine that assumes a priori knowledge of the data distribution. This poses a difficult challenge of choosing a metric that can be useful for generating necessary results for the particular task at hand.

Distance metric learning (or simply, metric learning) is used to automatically learn the task-specific distance metric from (weakly) supervised data. The learned metric transforms the sample space into a learned metric space where distances between samples are tailored specifically for the task at hand. One of the most widely used metric learning methods is the Large Margin Nearest Neighbor [23] or LMNN in short, which was designed specifically to improve the accuracy of the kNN algorithm.

The primary idea behind LMNN is to decrease the intra-class distances i.e. *pull* similarly labelled samples close and increase the inter-class distances i.e. *push* samples of different classes farther. Thus the LMNN algorithm tries to minimize a loss function that penalizes large intra-class distances on one hand and small inter-class distances on the other.

The concept of *target neighbors* is the crux of learning a metric using the LMNN algorithm. Suppose we have a set of data points $X = \{x_1, x_2..., x_N\}$ with corresponding labels $\{y_1, y_2..., y_N\}$ where $X \subset \mathbb{R}^d$. Given $x_i \in X$, a target neighbor of x_i is $x_j \in X$ for which $i \neq j$ and $y_i = y_j$. If x_j is a target neighbor of x_i, then it is represented as $j \rightsquigarrow i$ (not symmetric). The target

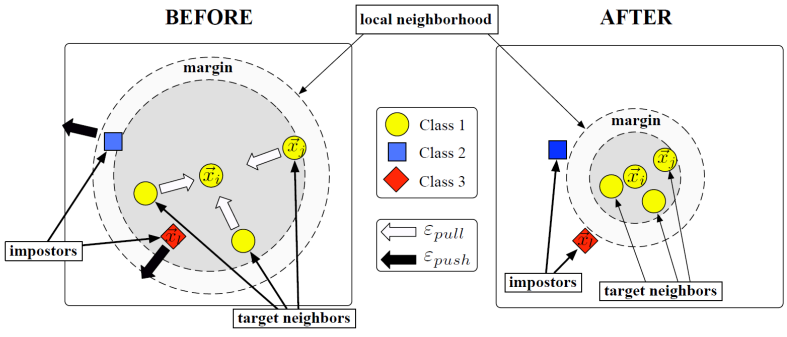


Figure 3.3: The push-pull concept of LMNN [23].

neighbors are fixed during the learning process and are decided by using any a priori knowledge available or by simply using the euclidean distance as a measure.

The k target neighbors can be assumed to form a circular perimeter of similarly labelled samples ideally that no differently labelled samples can invade. Any samples of a different class that is present within the perimeter are called *imposters*. The aim of optimizing the metric is to *push* these imposters away from the perimeter as shown in Fig. 3.3. Mathematically, the imposters ($x_l \in X$) can be represented as

$$||L(x_i - x_l)||^2 \leq ||L(x_i - x_j)||^2 + 1 \tag{3.1}$$

where L is a linear transformation over X and $x_i, x_j \in X$ with $j \rightsquigarrow i$. Metrics are computed by taking euclidean distance over L. The addition of the term **1** in Eq. 3.1, also known as the *margin* is required to ensure separation when all the target neighbors coincide at a single point giving 0 perimeter radius.

The loss function consists of two terms — one that *pulls* similarly labelled samples close and one that *pushes* dissimilar samples apart.

$$\varepsilon_{pull}(L) = \sum_{i=1}^{N} \sum_{j \rightsquigarrow i} ||L(x_i - x_j)||^2$$

$$\varepsilon_{push}(L) = \sum_{i=1}^{N} \sum_{j \rightsquigarrow i} \sum_{l=1}^{N} (1 - y_{il})[1 + ||L(x_i - x_j)||^2 - ||L(x_i - x_l)||^2]_+$$

where y_{il} is an indicator variable which is 1 iff $y_i = y_l$ and 0 otherwise. The factor $[z]_+ = max(z, 0)$ is the standard hinge loss. Samples that are going to contribute to ε_{push} are dissimilar ones ($y_i \neq y_l$) and for which the second factor is strictly positive i.e. x_l is an imposter and Eq. 3.1 holds. The gradient of ε_{pull} generates a pulling force that pulls the target neighbors close whereas the gradient of ε_{push} generates a pushing force that pushes the imposters away from the perimeter. The two losses are competing in nature and can

be combined to form a single loss using a weighting parameter $\mu \in [0, 1]$.

$$\varepsilon(L) = (1 - \mu)\varepsilon_{pull}(L) + \mu\varepsilon_{push}(L)$$

The choice of μ does not gravely affect the loss function and is usually chosen to be $1/2$. The above loss can be optimized as a semidefinite program (SDP) by reformulating it over positive semidefinite matrices as explained in [23].

3.3 Weibull Distribution

Named after Swedish mathematician Waloddi Weibull, the weibull distribution is a continuous probability distribution widely used in probability and statistics. It's wide usage can be attributed to it's versatility since it can assume the characteristics of many other types of distribution with the help of it's shape parameter β.

The probability density function (PDF) of a weibull distribution can be written as

$$\rho(x; \beta, \mu, \lambda) = \begin{cases} \frac{\beta}{\lambda}(\frac{x-\mu}{\lambda})^{\beta-1} e^{-(\frac{x-\mu}{\lambda})^{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\beta > 0$ is the shape parameter, $\mu \in (-\infty, +\infty)$ is the location

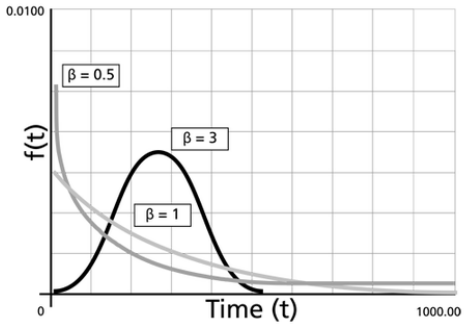


Figure 3.4: Effect of shape on weibull PDF. Source: reliawiki

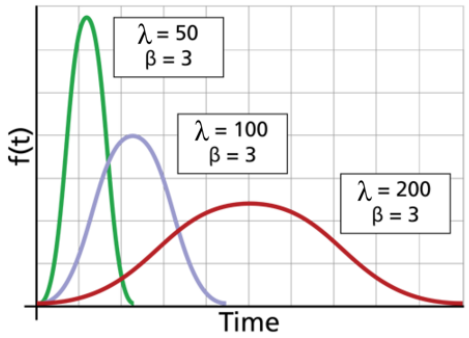


Figure 3.5: Effect of scale on weibull PDF. Source: reliawiki

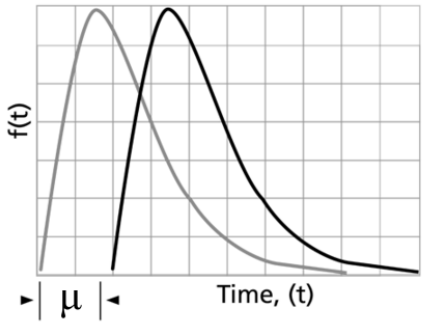


Figure 3.6: Effect of location on weibull PDF. Source: reliawiki

parameter and $\lambda > 0$ is the scale parameter. Due to the presence of three parameters in it's PDF viz. β, μ and λ, it is also known as a 3-parameter weibull distribution.

A 3-parameter weibull distribution is similar to a gaussian distribution when it's slop parameter $\beta = 1$ where the location parameter μ resembles the mean and the scale parameter λ resembles the variance.

A 2-parameter weibull distribution can be obtained by setting the location parameter $\mu = 0$. The PDF of such a distribution is

$$\rho(x;\beta,\lambda) = \begin{cases} \frac{\beta}{\lambda}(\frac{x}{\lambda})^{\beta-1} e^{-(\frac{x}{\lambda})^{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Chapter 4

Proposed Methodology

Problem Definition

Since we are detecting pedestrians in the wild [13], our gallery set G is generated dynamically with each frame encountered. As described in Sec. 1.5, for a standard re-ID in the wild problem, all the query/probe persons are not present in a single frame as a result of which the total the number of probe persons can be larger than the maximum size of G. Thus re-identifying pedestrians under such conditions is an open-set problem [55] where the probe set P is not guaranteed to be a subset of the gallery set (dynamic) G. Thus, a rejection mechanism is necessary to filter out the false alarms arising due to the absence of a probe person in G. Existing metric learning methods [23, 22, 56, 25] are unable to tackle this situation which has led us to develop a novel open-set metric learning system based on the concept of open-set recognition [57], LMNN [23] and weibull distribution. We name our method Open-Set LMNN or OS-LMNN.

It is to be noted that our solution proposes a one-shot re-ID method i.e. each frame is taken individually and independently for re-identification and temporal dependencies between frames are discarded. Thus it can be extended to solve image based re-ID systems having limited instances of a particular pedestrian. Our problem formulation can be broadly divided

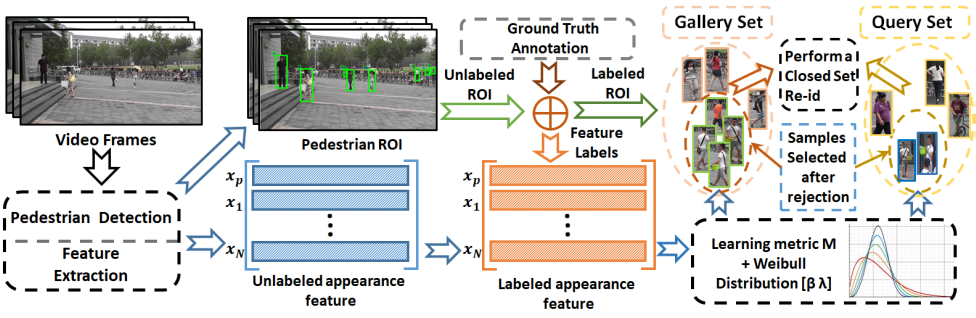


Figure 4.1: An overview of our proposed methodology

into three components as illustrated in Fig 4.1— (a) pedestrian detection followed by feature extraction, (b) joint optimization of the OS-LMNN with weibull parameters according to the proposed loss and (c) open to closed set (probe/gallery) conversion following weibull rejection before similarity ranking.

4.1 Pedestrian Detection

With the advent of deep learning for image recognition, several pedestrian detectors [27, 28, 51] have evolved featuring the"proposal + detector" approach. Among them, Mask R-CNN [52] yields the best localization results due to the use of a more accurate RoI Align compared to RoI pooling in it's previous versions and also due to the inclusion of Feature Pyramid Network (FPN) [53] in it's backbone for handling variations in scale (see Sec. 3.1). The backbone of our Mask R-CNN is ResNet101 + FPN as used in [58] which is available at https://github.com/matterport/Mask_RCNN.

The Mask R-CNN is pre-trained on ImageNet and is fine-tuned on the PRW [13] training dataset as a 2-class recognition model i.e. whether an image contains a pedestrian or not. As shown in Fig. 4.2, only pedestrians (labelled as *person*) are extracted from the network and objects having

Figure 4.2: Mask R-CNN results on a frame from the PRW dataset

other labels are discarded. Our problem just requires the use of bounding boxes to extract the pedestrian RoIs from each frame and hence the segmentation mask and confidence scores that are also generated by the network are not used. Feature extraction is a key part of this section since we need to represent each pedestrian by their ID-discriminative embedding (IDE) as mentioned in [13]. In general for feature extraction, the last fully-connected (FC) layer of the network is used as the feature vector which is a compact representation of the learnt features of each object. But since our network is trained to classify between an image having a pedestrian or not, all the pedestrians are detected under a single label i.e. *person*. Thus discriminating between different instances (pedestrians) of the same class (person) is not feasible since their feature vectors (extracted from the last FC layer) have a strong similarity for valid classification. This will adversely affect the later stages of metric learning.

This issue of learning discriminative feature representations for differ-

ent pedestrians is solved by using traditional descriptors on top of R-CNN based detectors. Traditional descriptors are preferred over deep models because they are easy to train for pedestrian IDs having low instances. In this work, we have used four feature descriptors that are Bag-of-Words (BoW) vector [59], HistLBP [60], Local Maximal Occurrence (LOMO) [56] and gBiCov [61]. The choice of descriptors were based on previous literature for justifiable comparison with the baseline. All of them were trained on the PRW training dataset with 483 classes representing the 482 pedestrian IDs + 1 ID for unknown pedestrians ("-2") .

4.2 Open-Set Metric Learning (OSML)

Previous works on open-set person re-ID have followed the idea of rejecting the probe samples that are unlikely to be present in the gallery thus converting the open-set problem to a close-set one. A trivial way to do this is by the use of a heuristic threshold over similarity measures [55]. Some recent works [57, 62] based on the concept of Extreme Value Theorem (EVT) have shown that a learnt weibull distribution (see Sec. 3.3) can better represent unlikely samples that fall at the tail of the distribution. Hence, such samples can be easily separated with the use of a smaller threshold over probability. In our proposed approach, weibull distribution parameters corresponding to each pedestrian ID are learnt along with a Mahalanobis distance metric by alternatively optimizing a regularized error function.

Suppose x_i and x_j are the feature representations of a probe sample $i \in P$ and a gallery sample $j \in G$ respectively. The distance $D_{ij}^{\mathbf{M}}$ between the samples i and j is represented using the Mahalanobis distance metric $\mathbf{M}$ as

$$D_{ij}^{\mathbf{M}} = (x_i - x_j)^T \mathbf{M} (x_i - x_j) \tag{4.1}$$

The metric $\mathbf{M}$ in the above equation can be learnt by the LMNN approach

by optimizing the following error (loss) function (see Sec. 3.2)

$$\varepsilon(\mathbf{M}) = (1-\mu)\sum_{i \rightsquigarrow j} D_{ij}^{\mathbf{M}} + \mu \sum_{i,j \rightsquigarrow i} \sum_{k} [\alpha + D_{ij}^{\mathbf{M}} - D_{ik}^{\mathbf{M}}]_{+} \tag{4.2}$$

where the first (pull) error term corresponds to distance between similar pairs $(i, j; i \rightsquigarrow j)$ and the second (push) error term corresponds to distance margin (α) that an imposter sample (k) can intrude a true sample (j) with respect to an anchor (i). All such combinations of anchor (i), true (j) and imposter (k) samples form a valid triplet and $[z]_{+} = max(z, 0)$ is the standard hinge loss. Here, $\mu \in [0, 1]$ is the weighting parameter that balances the push and pull factors which is usually fixed at a particular value. But in our OS-LMNN approach, we dynamically adjust this weight to better separate the imposters from the ID distribution of true samples which is explained mathematically in Eq. 4.5. After learning a weibull distribution [57], the dynamic weight is computed based on the weibull parameters. Here, a 2-parameter (β, λ) weibull distribution suffices our purpose since the ID distributions can be assumed to have a fixed location in the learnt metric space. The standard probability density function (PDF) of a 2-parameter weibull distribution can be written as

$$\rho(x; \beta, \lambda) = \begin{cases} \frac{\beta}{\lambda}(\frac{x}{\lambda})^{\beta-1} e^{-(\frac{x}{\lambda})^{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{4.3}$$

where $\beta > 0$ and $\lambda > 0$ are respectively the shape and scale parameters of the weibull distribution [57]. The cumulative distribution function (CDF) of the 2-parameter weibull distribution can be written as

$$F(x;\ \beta, \lambda) = [1 - e^{-(x/\lambda)^{\beta}}] \tag{4.4}$$

where $F \in [0, 1]$ is a bounded monotonically increasing function. Utilizing these characteristics of a weibull CDF we jointly learn the Mahalanobis met-

ric along with the weibull distribution based on our proposed error function as follows.

$$\begin{aligned}\varepsilon(\mathbf{M},\beta,\lambda) = \sum_{i,j\rightsquigarrow i}\sum_{k}\Big\{\Big(\frac{\omega_{ki}}{1+\omega_{ki}}\Big)\cdot D_{ij}^{\mathbf{M}} \\ + \Big(\frac{1}{1+\omega_{ki}}\Big)\cdot[\alpha + D_{ij}^{\mathbf{M}} - D_{ik}^{\mathbf{M}}]_{+}\Big\}\end{aligned} \tag{4.5}$$

where $\omega_{ki} = F(D_{k\mu_i}^{M};\beta,\lambda)$ (see Eq. 4.4). x_{μ_i} is the mean feature vector of samples of person i belonging to same ID and x_k is the feature vector for an imposter sample. Based on the weibull CDF, smaller distance between dissimilar pairs decreases ω_{ki} thereby increasing the push factor weight $1/(1+\omega_{ki})$ with respect to the pull one and vice-versa. weibull parameters $\mathbf{w} = [\beta\ \lambda]$ learnt based on Eq. 4.5 will saturate to an arbitrary high value. In order to restrict such saturation, we regularize the above error function as:

$$\mathbf{M}^*,\beta^*,\lambda^* = \underset{M,\beta,\lambda}{\operatorname{argmin}}[\varepsilon(\mathbf{M},\beta,\lambda) + \gamma\cdot\mathcal{R}(\beta,\lambda)] \tag{4.6}$$

Here $\mathcal{R}(\beta,\lambda) = \frac{1}{2}\mathcal{N}\cdot(\beta+\lambda)$ is a regularization term where $\mathcal{N}$ is the total number of valid triplets. We have adopted L-BFGS-B [63] optimizer to minimize the objective function $J = \varepsilon(\mathbf{M},\beta,\lambda) + \gamma\cdot\mathcal{R}(\beta,\lambda)$ (Eq. 4.6) via alternatively fixing $\mathbf{M}$ and $\mathbf{w}$ at each iteration. The gradient of J with respect to $\mathbf{M}$ can be computed as:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{M}} = \sum_{i,j\rightsquigarrow i}\sum_{k}\Big\{\Big(\frac{\omega_{ki}}{1+\omega_{ki}}\Big)\cdot\mathbf{C}_{ij} + \Big(\frac{1}{1+\omega_{ki}}\Big)\cdot(\mathbf{C}_{ij}-\mathbf{C}_{ik}) \\ + \Big(\frac{1}{1+\omega_{ki}}\Big)^2\cdot\rho(D_{k\mu_j}^{\mathbf{M}})\cdot\big(D_{ij}^{\mathbf{M}} - [\alpha + D_{ij}^{\mathbf{M}} - D_{ik}^{\mathbf{M}}]_{+}\big)\cdot\mathbf{C}_{\mu_i k}\Big\}\end{aligned} \tag{4.7}$$

where $\mathbf{C}_{ij} = (x_i - x_j)(x_i - x_j)^T$ is the outer product of samples x_i and x_j and $\mathbf{C}_{\mu_i k} = (x_{\mu_i} - x_k)(x_{\mu_i} - x_k)^T$ is the outer product of x_{μ_i} and x_k. The

gradient of J with respect to $\mathbf{w}$ is:

$$\frac{\partial J}{\partial \mathbf{w}} = \begin{cases} \frac{1}{(1+\omega_{ki})^2} \cdot \frac{\partial \omega_{ji}}{\partial \mathbf{w}} \sum_{i,j \rightsquigarrow i,k} \sum D_{ij}^{\mathbf{M}^*} + \frac{1}{2}\gamma \cdot \mathcal{N} & \varepsilon_{push} \leq 0 \\ \frac{1}{(1+\omega_{ki})^2} \cdot \frac{\partial \omega_{ji}}{\partial \mathbf{w}} \sum_{i,j \rightsquigarrow i,k} \sum [D_{ik}^{\mathbf{M}^*} - \alpha]_+ + \frac{1}{2}\gamma \cdot \mathcal{N} & \varepsilon_{push} \geq 0 \end{cases} \tag{4.8}$$

where $\varepsilon_{push} = [\alpha + D_{ij}^{\mathbf{M}} - D_{ik}^{\mathbf{M}}]_+$ and

$$\frac{\partial \omega_{ji}}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial \omega_{ji}}{\partial \beta} \\ \frac{\partial \omega_{ji}}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} -\frac{\beta}{\lambda} \cdot \left(\frac{D_{k\mu_j}^M}{\lambda}\right)^{\beta} \cdot e^{-(D_{k\mu_j}^M/\lambda)^{\beta}} \\ ln\left(\frac{D_{k\mu_j}^M}{\lambda}\right) \cdot \left(\frac{D_{k\mu_j}^M}{\lambda}\right)^{\beta} \cdot e^{-(D_{k\mu_j}^M/\lambda)^{\beta}} \end{bmatrix} \tag{4.9}$$

4.3 Weibull Rejection

As described in Sec. 4.2, a weibull distribution (parameterized by $\mathbf{w} = [\beta\ \lambda]$) is learnt along with a Mahalanobis metric $\mathbf{M}$ over the extracted pedestrian features. The weibull distribution gives a compact abating probability (CAP) [57] model of each ID distribution that is uniquely identified by the point x_{μ_i} in the feature space. The learnt distribution assigns a low similarity (probability) value to dissimilar pairs in the new learnt metric space (both the distribution and the metric are learnt alternatively). Thus given a frame i.e. given a dynamically generated gallery set G, every probe sample is checked with the gallery ID distributions and assigned a similarity value with each of them. A gallery ID rejects a probe sample if it's similarity is less than a particular assigned threshold (weibull rejection). If the probe sample is rejected by all the gallery samples in G, then the probe is inferred to be absent in that frame. Similarity rankings are then performed only with the remaining probe samples i.e. probe samples that has not been rejected by the gallery G (open to close-set conversion). The gallery ID having the highest similarity with the probe sample is chosen to be the ID of the probe thus completing re-identification in the wild.

Chapter 5

Experiments

In this chapter, we describe the dataset used for evaluating our model, the implementation details, the evaluation measures used for comparison and finally we present a detailed comparison with the baseline and a combination of different feature extractors and metric learning methods.

5.1 Dataset Description

For evaluating our model, we have used the PRW dataset [13]. It consists of 11,816 frames captured from 6 different views at 25 fps with 5 views having a resolution of 1080×1920 and the remaining view having a resolution of 576×720. Pedestrians are annotated at every 25^{th} frame i.e. 1 annotated frame per second. Out of a total of 43,100 annotated pedestrians, 34,304 are assigned ID ranging from "1" to "932" and the rest ambiguous ones are assigned an ID of "-2" which refers to *unknown*. Most re-ID datasets presents only cropped pedestrian RoIs (annotated bounding boxes) instead of the raw frames which restricts our model's evaluation for *re-identifying person in the wild* i.e. simultaneous detection and re-identification. As a result, we have evaluated and presented our performance on the PRW dataset only where full frames sampled at 1 fps directly from the camera feed are available.

5.2 Evaluation Measures

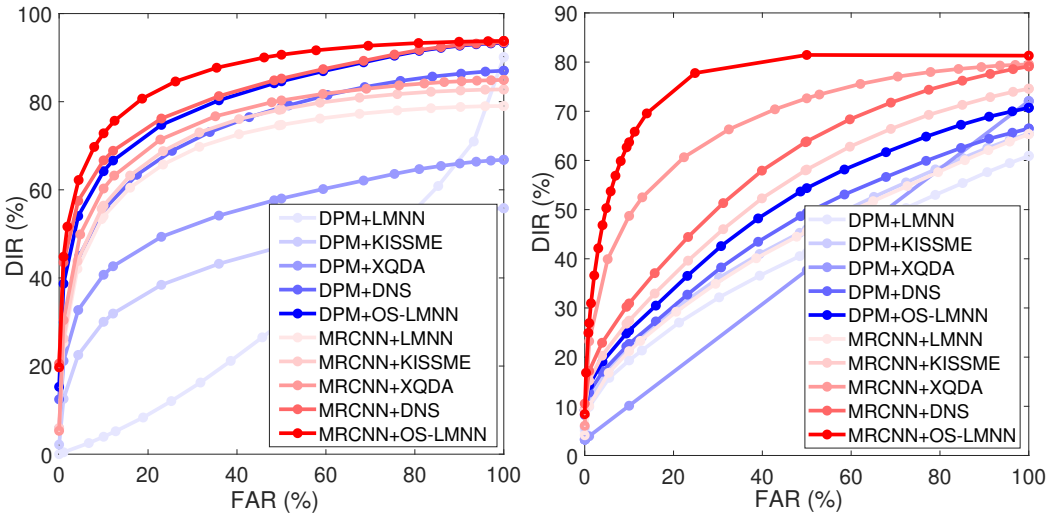
For evaluating our OS-LMNN model, a modified version of the two measures as given in [55], viz. Detection and Identification Rate (DIR) and False Accept Rate (FAR) are used. Usually, ROC is obtained by varying a threshold τ over distance based similarity. We instead vary τ over Weibull distribution (ρ) as:

$$\begin{aligned} DIR(\tau, k) &= \frac{|\{p : p \in P_G,\ rank(p) \leq k,\ \rho(D^{\mathbf{M}}_{pg}) \geq \tau\}|}{|P_G|} \\ FAR(\tau) &= \frac{|\{p : p \in P_N,\ and\ \rho(D^{\mathbf{M}}_{pg}) \geq \tau\}|}{|P_N|} \end{aligned} \tag{5.1}$$

where P_G, P_N are the two probe sets, G is the gallery set and $g \in G$. Here G consists of persons common to set P_G but not in set P_N. Cumulative Matching Characteristic (CMC) curves are then constructed from *rank-k* recognition rate. Please note that DIR vs. FAR becomes CMC, when FAR = 100%. For having fair comparisons with other metric learning methods, we have normalized the distance between any two samples to $[0, 1]$ based on the maximum pairwise distance before using Eq. 5.1.

5.3 Implementation Details

All the experimentation and implementations are done in *Python3* using open-source libraries like numpy, scipy, sklearn, etc. For the pedestrian detector, Mask R-CNN [52] pre-trained on ImageNet was fine-tuned (transfer learning) on the PRW dataset (see Sec. 4.1) using keras over a tensorflow backend. Our framework has two hyper-parameters, namely, (a) margin α (Eq. 4.5) and (b) regularization constant γ (Eq. 4.6) which are experimentally set to 25 and 0.5 respectively.



(a) With BoW [59] descriptor

(b) With HistLBP [60] descriptor

(c) With gBiCov [61] descriptor

(d) With LOMO [56] descriptor

Figure 5.1: ROC Curve with DIR vs FAR comparison at rank-1 recognition rate for different feature descriptors.

5.4 Performance Comparison

The performance of our model is compared with four feature descriptors and four metric learning methods for each of the two state-of-the-art pedestrian detectors, viz. DPM [64] and Mask R-CNN [52]. The feature descriptors used are Bag-of-Words (BoW) vector [59], HistLBP [60], LOMO [56] and gBiCov [61]. The metric learning methods used for comparison are LMNN [23], KISSME [22], DNS [25] and XQDA [56]. It is to be noted that currently, no method with or without deep learning, is available that inherently deals with the problem of open-set re-ID in the wild.

Table 5.1: DIR vs. varying FAR for Rank-1 scores with the DPM detector. (Best values are shown in **bold**.)

Detector	Feature	Recognizer	FAR(%)				AUC (%)
			1	10	50	100	
DPM [64]	HistLBP [60]	LMNN [23]	9.89	19.34	41.21	60.92	39.64
		KISSME [22]	11.17	21.83	46.01	65.45	43.88
		DNS [25]	12.79	22.70	49.34	66.50	45.98
		XQDA [56]	3.92	10.12	37.65	**72.06**	37.65
		OS-LMNN (Ours)	**14.31**	**25.40**	**54.37**	70.71	**50.26**
	LOMO [56]	LMNN [23]	12.64	28.69	61.56	65.58	53.89
		KISSME [22]	15.95	40.60	63.11	68.98	58.34
		DNS [25]	23.12	43.72	70.87	77.58	65.22
		XQDA [56]	21.97	41.54	67.33	73.70	61.96
		OS-LMNN (Ours)	**30.04**	**57.63**	**81.42**	**87.61**	**76.30**
	BOW [59]	LMNN [23]	0.43	3.96	29.84	90.11	32.85
		KISSME [22]	12.56	30.04	47.07	55.82	44.35
		DNS [25]	30.03	55.30	78.71	87.03	74.19
		XQDA [56]	21.17	40.72	58.02	66.82	55.16
		OS-LMNN (Ours)	**38.69**	**64.23**	**84.62**	**93.34**	**80.83**
	gBiCov [61]	LMNN [23]	10.12	23.41	54.70	62.31	48.24
		KISSME [22]	15.95	40.60	63.11	68.98	58.34
		DNS [25]	23.12	43.72	70.87	**77.58**	**65.22**
		XQDA [56]	17.57	33.22	53.86	58.96	49.57
		OS-LMNN (Ours)	**24.03**	**46.10**	**65.14**	70.09	61.40

We first perform an exhaustive analysis with DPM as the pedestrian detector and the results are shown in Tab. 5.1. The bounding boxes detected by DPM have many false positives as compared to Mask R-CNN. Yet our proposed metric learning model (OS-LMNN) has performed well with respect

to the other metric learning methods. As can be seen from the Tab. 5.1, OS-LMNN has outperformed various combinations of different feature descriptors and different metric learning methods by attaining a maximum DIR value of 38.69% at 1% FAR. The best competitor among other methods is BoW [59] + DNS [25] with a DIR of 30.03% at 1% FAR. This combination is also slightly lagging behind our 2^{nd} best combination LOMO [56] + OS-LMNN (DIR = 30.04% at 1% FAR). Also at other values of FAR, specifically, at 10%, 50% and 100% we have secured highest performance with DIR of 64.23%, 84.62% and 93.34% respectively. Some metric learning methods drastically under-performed with DPM detectors as the influence of false positives has overwhelmed the genuine pairs.

Table 5.2: DIR vs. varying FAR for Rank-1 scores with Mask R-CNN detector. (Best values are shown in **bold**.)

Detector	Feature	Recognizer	FAR(%)				AUC (%)
			1	10	50	100	
Mask R-CNN [52]	HistLBP [60]	LMNN [23]	10.08	21.23	45.54	65.36	43.23
		KISSME [22]	15.21	27.42	58.14	74.59	53.68
		DNS [25]	17.19	30.92	63.87	79.13	58.54
		XQDA [56]	23.23	48.72	72.64	79.54	67.61
		OS-LMNN (Ours)	**26.92**	**63.71**	**81.45**	**81.51**	**76.11**
	LOMO [56]	LMNN [23]	24.81	49.45	70.79	75.9	65.99
		KISSME [22]	26.63	57.95	80.3	84.73	74.86
		DNS [25]	**31.96**	59.41	87.68	93.04	81.29
		XQDA [56]	28.03	61	84.53	89.19	78.80
		OS-LMNN (Ours)	31.65	**72.7**	**92.65**	**93.31**	**87.15**
	BOW [59]	LMNN [23]	25.3	54	74.74	79.02	69.83
		KISSME [22]	26.25	56.41	78.27	82.78	73.10
		DNS [25]	43.51	66.68	85.25	93.48	81.85
		XQDA [56]	30.41	60.29	80.28	84.95	75.49
		OS-LMNN (Ours)	**44.82**	**72.82**	**90.66**	**93.8**	**86.23**
	gBiCov [61]	LMNN [23]	11.62	21.93	56.01	71.54	49.17
		KISSME [22]	18.05	39.39	57.48	62.18	53.58
		DNS [25]	24.88	41.9	63.69	69.06	59.13
		XQDA [56]	24.06	41.72	67.57	80.6	63.50
		OS-LMNN (Ours)	**28.04**	**51.49**	**79.4**	**90.61**	**74.24**

Mask R-CNN based detections are more accurate and it further increases the overall efficiency of our OS-LMNN model (see Tab. 5.2). We have achieved a 6% performance boost with DIR of 44.82% for our best combina-

tion OS-LMNN + BoW at 1% FAR. Among other methods DNS + BoW has performed best but with a lower DIR of 43.51% at 1% FAR. The improved performance of DNS is due to the incorporation of non-linearity achieved through kernels. Still, for several combinations, our proposed OS-LMNN, in-spite of its linear nature, has clearly outperformed DNS.

The impact of jointly learning the Mahalanobis metric using LMNN with the weibull distribution is evident from the ROC plots of Fig. 5.1. For all cases, the vanilla LMNN model has performed miserably low as compared to our model. For the DPM detector, the best descriptor combination with LMNN is LOMO with a DIR of 12.64% which is pretty low as compared to ours (30.04%) at 1% FAR. Though the performance of LMNN has improved for Mask R-CNN detector (DIR = 24.81%), it still considerably lags with respect to our OS-LMNN. This clearly shows the impact of weibull rejection based metric learning especially when open-set is considered.

Chapter 6

Conclusion

In this paper, we have proposed a new metric learning model especially for performing open-set re-ID in wild. We have introduced the concept of weibull rejection by learning ID distributions via weibull and setting a similarity threshold to reject the probe samples absent in the gallery thus converting the open-set problem to a close-set. Our model can be further improved by introducing non-linearity through kernels that can better represent the metric space for complex inputs. Instead of using a deep model for pedestrian detection and a statistical machine learning model for re-identification, in the future, we plan on extending our open-set metric learning framework to fully deep architectures that can be trained end-to-end.

Bibliography

[1] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and vision computing*, vol. 32, no. 4, pp. 270–286, 2014.

[2] T. Huang and S. Russell, "Object identification in a bayesian context," in *IJCAI*, vol. 97, pp. 1276–1282, 1997.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[4] W. Zajdel, Z. Zivkovic, and B. Krose, "Keeping track of humans: Have i seen this person before?," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 2081–2086, IEEE, 2005.

[5] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1528–1535, IEEE, 2006.

[6] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *2010 20th International Conference on Pattern Recognition*, pp. 1413–1416, IEEE, 2010.

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local fea-

tures," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2360–2367, IEEE, 2010.

[8] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *2014 22nd International Conference on Pattern Recognition*, pp. 34–39, IEEE, 2014.

[9] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159, 2014.

[10] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3908–3916, 2015.

[11] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 737–753, 2018.

[12] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 937–940, 2014.

[13] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1367–1376, 2017.

[14] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*, pp. 262–275, Springer, 2008.

[15] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586–3593, 2013.

[16] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 144–151, 2014.

[17] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European conference on computer vision*, pp. 330–345, Springer, 2014.

[18] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, pp. 3739–3747, 2015.

[19] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4184–4193, 2015.

[20] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.

[21] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, p. 4, 2006.

[22] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 2288–2295, IEEE, 2012.

[23] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[24] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3685–3693, 2015.

[25] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1239–1248, 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[28] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[29] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, Lille, 2015.

[30] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[31] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European conference on computer vision*, pp. 135–153, Springer, 2016.

[32] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[33] A. Sikdar and A. S. Chowdhury, "Scale-invariant batch-adaptive residual learning for person re-identification," *Pattern Recognition Letters*, vol. 129, pp. 279–286, 2020.

[34] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–8, IEEE, 2016.

[35] L. Wu, C. Shen, and A. Van Den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

[36] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*, pp. 91–102, Springer, 2011.

[37] A. Bedagkar-Gala and S. K. Shah, "Part-based spatio-temporal model for multi-person re-identification," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1908–1915, 2012.

[38] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 33–40, 2015.

[39] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1354–1362, 2016.

[40] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European conference on computer vision*, pp. 688–703, Springer, 2014.

[41] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3810–3818, 2015.

[42] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang, "Temporally aligned pooling representation for video-based person re-identification," in *2016 IEEE international conference on image processing (ICIP)*, pp. 4284–4288, IEEE, 2016.

[43] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5683–5695, 2018.

[44] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*, pp. 868–884, Springer, 2016.

[45] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1325–1334, 2016.

[46] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*, pp. 701–716, Springer, 2016.

[47] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2016.

[48] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2613–2622, 2016.

[49] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[50] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[53] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[55] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li, "Open-set person re-identification," *arXiv preprint arXiv:1408.0872*, 2014.

[56] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2197–2206, 2015.

[57] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014.

[58] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow." https://github.com/matterport/Mask_RCNN, 2017.

[59] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

[60] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European conference on computer vision*, pp. 1–16, Springer, 2014.

[61] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.

[62] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 762–768, 2017.

[63] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[64] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.