ACKNOWLEDGEMENT

The present thesis work is an outcome of hard work and contribution of many well-wishers, whose support, help, suggestions and guidance has resulted in timely completion of project.

I would like to express my deepest gratitude and thanks to my supervisor **Dr. Dinesh K. Vishwakarma**(Assistant Professor), Department of Electronics and Communication Engineering, Delhi Technological University. His vital support, quality guidance, precious knowledge and constant motivation from the initial to the final level of this work enabled me to develop an understanding of the subject and helped me in the timely success of my project and submission of this thesis.I am highly thankful to him for guiding me in this project.

I am also extremely thankful to **Dr. S. Indu**, Head of the Department of Electronics and Communication Engineering, Delhi Technological University, for the support provided by her during the entire duration of degree course and for providing all the facilities to carry out quality project work.

I also wish to express my appreciation to the classmates as well as staff at Department of Electronics and Communication of Delhi Technological University for their goodwill and support that helped me a lot, in successful completion of this project.

Finally, I want to thank my parents, for inculcating good ethos, as a result of which I am able to do my post-graduation from such an esteemed institution. I would thank my friends for believing in my abilities and for always showering their invaluable support and constant encouragement.

Sakshi upadhyay
2K15/SPD/14
M.Tech (SPDD)
Dept. of ECE, DTU, Delhi

ABSTRACT

The field of surveillance and forensics research is currently shifting focus and is now showing an ever increasing interest in the task of people re-identification. It is a fundamental task in the automated surveillance system essential to track a person in multi-camera setting. Re-identification (Re-ID) can be defined as a process of identifying the resemblances of a set of probe images or a single probe image representing a single person from a set of gallery images of people taken from the same or different cameras placed at different locations.

However, established identification techniques being used presently face many difficulties and shortcomings. Traditional surveillance cameras provide low resolution images and thus state of the art face recognition and iris recognition algorithms cannot be easily applied to surveillance videos and images as people are required to face the camera at a close range. The different lighting environment inherited by each camera scene and the strong variations in illumination induce large changes in their appearance of a person walking through the scene. In addition, people images are occluded by other passers-by or objects in the scene making people detection further difficult to achieve.

So to address the challenges in person re-identification problem, major contributions are being made to design robust feature representations and good discriminant metrics to evaluate the similarity between two person images effectively. Recently, it seems that more researches have been made in metric learning. So in this work potentials of feature design are emphasized. A novel and efficient person descriptor is proposed by utilizing knowledge on dense sampling of low-level statistics. We simply model a multi-layer representation of pixel features using multiple Gaussian distributions. More specifically, first we have created a pixel feature representation using color and texture information, then we have extracted local patch Gaussians inside overlapping regions. Further to describe these regions we have again used a Gaussian model on the local patch Gaussians. Thus we are able to use the discriminative properties of mean and covariance together along with considering the local structures in the image while globally analyzing them.

For metric learning we have used a discriminant subspace and kernel learning method described in *Liao et. al (2015)*, i.e. XQDA. It learns a discriminant low dimensional subspace and a QDA metric on the projected subspace, simultaneously.

The proposed descriptor and metric is compared on benchmark datasets with current state-of-art-methods and has proven to give efficient and robust results. It exhibits remarkably high performance which outperforms some of the state-of-the-art descriptors for person re-identification. We have then considered it as a retrieval or recognition problem with the expectation that the n-highest ranked matches in the gallery will provide an identity for the unknown person, thereby identifying the probe.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Nowadays, widespread networks of cameras are being used in various public places like railway stations, airport, hospitals, office buildings, shopping malls, college campuses and streets. They cover large areas and have non-overlapping view-points and thus provide huge amount of data encoded well within the images/videos. This data can be utilized effectively for public safety applications, including surveillance for threat detection, detection of unusual events, monitoring of elderly people and patients in hospital, customer tracking in stores, etc. Manual monitoring of this data is cumbersome, time consuming and prone to errors thus reducing the efficiency of human surveillance system. Therefore, researchers are working towards automating this procedure using computer vision system. There has been impressive progress in pattern recognition and machine learning techniques recently which has consequently given more efficient automated surveillance system and improved the scope of usage in forensics industries.

Understanding of a surveillance scene through computer vision requires the ability to track people across multiple cameras, perform crowd movement analysis and activity detection. Person Re-Identification is an elementary task in this automated surveillance system essential to track a person in multi-camera setting. It is referred as a process of identifying the resemblances of a group of probe images or a single probe image representing a single person from the other group called gallery images of people captured from same or different cameras placed at different locations. In simple words, if we feed an image/video of a person taken from one camera to the system, the system tries to identify the instances of the same person from images/videos taken from disjoint cameras. Depending on the number of available images per individual, two types of frame settings are considered in person re-identification: (a) single-shot (1) (2), if only one frame is captured per individual for both probe and gallery sets; and (b) multiple-shot (1) (2), for multiple frames per individual gathered over time using multiple views of the subjects. Besides surveillance it has applications in robotics, multimedia, and many more popular utilities like automated photo tagging or photo browsing.

People Re-identification can also be defined as the task of assigning the same identifier to all the instances of the same object or, more specifically, of the same person, by utilizing some kind of visual properties that have been captured and extracted from an image or a video. As humans, we do this task all the time without much effort. Our eyes and brain is trained to detect, localize, identify and later re-identify objects and people in the real world. Technically speaking, it can be broken down into three modules, i.e., *person detection, person tracking, and person retrieval*. It is generally believed that the first two modules are independent computer vision tasks, so most re-ID works focus on the last module, i.e., person retrieval.

The person re-identification challenges mainly have three broad steps. First, it is determined that which part of the image is the area of interest and thus needs to be segmented before its analysis and comparison to find the matches. Second, invariant discriminative signatures of each individual are then constructed for analogous parts to be compared. Third, a suitable and discriminative metric must be applied to match these signatures to identify the correct instances of a person. In this text, we assume that the appearance of the person, specifically the clothes and the rare objects (like bagpacks, purse, etc.), does not change while it appears in different cameras in short span of time disregarding the case of considering images taken days or months apart. This is called short-period Re-Id. Then the simplest and most obvious descriptors considered are appearance based descriptors of a person, utilizing features like color, shape and texture. But it is not an easy job to tackle the several challenges offered by the problem.

1.1 CHALLENGES

Challenges faced by this task of re-identification increases the difficulty level as compared to just the identification task. The challenges differ in different scenarios, however, all applications face some common problems.

- Illumination Variations - Illumination conditions may vary largely due to difference in camera settings and as well as in the same camera, in different phases of day and night, due to changes in the open environment conditions. Brightness levels in outdoor situations may consistently differ amid different time durations as climatic conditions change and light of the sun fluctuates. Then, again the lighting state in indoors may deviate between two cameras because of various sorts of lights glowing (neon, tungsten).This results in significant variations in the appearance of a person across disjoint camaras and during different time periods.
- Pose and View-Angle Variations - The relative pose of a person, as he walks across the camera views in a camera network, varies with the walking path and direction of that person, and also depends on the camera view angle or scale changes inherent in a multi-camera setting. This may create problems for gait based methods focused on the gait of the

moving person as it constantly varies across different view-angles and poses. This is a serious issue as it largely diminishes the recognition rates.

- Occlusion - Partial or complete occlusion is another issue and may be caused by objects, clothing accessories and other humans in the scene. It can lead to the failure of segmentation algorithms which try to separate target person from the rest of the scene.
- Changes in color response – One camera usually differs from others in its color sensitivities and thus sees the same colors slightly in difference to others which can sometimes largely affect the person's appearance. Color response of the sensors may likewise deviate succumbing to open conditions outside and due to the programmed color normalization that frequently happens in-camera.
- Low Resolution images - Moreover, due to relatively low resolution of the traditional surveillance cameras which capture very little or none of the facial properties, images of two different individuals may appear closer than the true image pairs, thus preventing the use of state of the art biometric and soft-biometric approaches.
- However, another challenge is less studied in existing work, which is "cross dataset person re-identification". In practical systems, a large dataset is collected first and then a model is trained on them. Then this model is superimposed on other datasets or videos for person re-identification. We call the training datasets as source domain and test datasets as target domain. The source and target datasets are totally dissimilar, because they are usually captured by different cameras under different environments, i.e., have different probability distributions. A practical person re-identification algorithm is expected to have good adaptation to the dataset changes. Therefore, cross dataset person re-identification is an important rule to analyze the performance of algorithms in practice.

So comparing person descriptors is challenging due to the uncertainty attributed by the possible lack of prior known spatio-temporal relationships between cameras. Additionally, appearance of the same person can differ drastically due to changes in other objects like bags, unzipped jackets across front and back views, etc. when at the same time appearance of different people might be rather similar. This implies that within class variations can be larger than inter-class variations.

Moreover, even if the person's descriptors may be captured effectively, matching them across cameras in the presence of large number of people observed is non-trivial. Comparing these descriptors across large number of potential candidates is a hard task as the descriptors are captured in different locations, time instants, and over different durations. Complexity of the matching process further increases with increase in the number of candidates leading to loss of descriptor non-ambiguity, increasing the possibility of matching errors. It also requires intensive memory and high computation capabilities.

1.2 ADDRESSING ISSUES

Person Re-ID system consists of mainly three stages as shown in Figure 1. Jointly they lead to two elementary problems which become the center of our focus in handling the task, i.e. feature representation and metric learning. An efficient feature representation is desired to have robustness to illumination and camera viewpoint changes, and a discriminant metric is required for matching given people images. Most of the trouble is taken along these two subjects to address the challenge of person re-identification. Most of the approaches followed in re-ID acquire appearance-based features that are viewpoint quasi-invariant (3) (4) (1) (5), such as color and texture models. However, different approaches vary greatly in the quantity and structure of features used and it becomes difficult to fairly evaluate how they affect the performance..

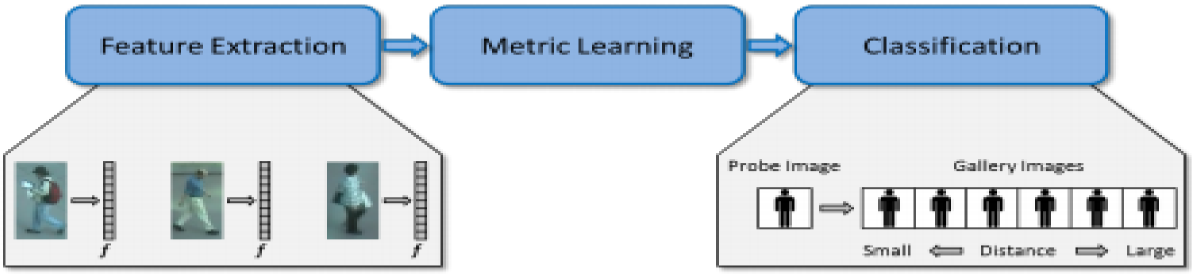


Fig.1: Person Re-Identification System.

Using fixed simple metrics like Euclidean distance on these basic feature types result in poor matching results since they consider all kinds of features equally important neglecting that some features are weak due to large variations in pose and illumination and limited training data. Thus, researchers have started designing classifiers that learn advanced metrics (4) (6) (7), which compel the features of the true individual to be at lesser distance than that from different individual. Yet, most algorithm results are still low, slightly above 30% for the nearest match.

Now coming to appearance based descriptors that exploit the appearance information of person’s clothing, specifically color and texture, robust models are tried to be constructed that can tackle intra-person variations and also provide good discrimination from other persons. Most works used these basic features (i) Color space values (1), usually embedded in histograms, (ii) object figure information, e.g. HOG descriptor (8), (iii) imbibed texture, often extracted by Gabor filters (9) (5), Schmid filters and differential filters (5) (iv) interest point descriptors, e.g. SURF and SIFT.

To address the low resolution issue and varying poses, the color information like color histograms of desired color channels and color name descriptors (10) come to the rescue and is able to discern two different person images. But for this an assumption is made that person clothes are not changed in the process. Some texture descriptors have also been successful in this objective namely, Maximally Stable Color Regions (MSCR) (11), Local Binary Patterns (LBP) and 21 texture filters (8 Gabor filters and 13 Schmid filters). But color information have shown better outcomes in comparison to texture data as in most of the scenes resolution achieved is low. However, they have their own drawbacks, as illumination variations and camera parameters change them and also cannot sufficiently differentiate different persons of similar color. Additionally, the color histogram discards any information regarding the spatial structures and texture of person data in images. These downsides confine the utilization of the color based methods in the person re-identification frameworks. So the performance of feature representation by color histograms methods is likewise not attractive.

Since those rudimentary features (color, shape, texture, etc.) captured diverse domains of the data imbibed in these images, many a times they are fused together to serve as a high quality signature mark. Some of those effective approaches are mentioned here: (12) combine the 8 color features and 21 texture filters (Gabor and differential filters). (13) (14) fuse together Maximally Stable Color Regions (MSCR) and weighted Color Histograms (wHSV), achieving outstanding results, and he ensemble of local features (ELF) (5), SDALF (14), kBiCov (15), fisher vectors (LDFV) (16), salience match. (17), and mid-level filter (18), all of these do some kind of fusion. These handcrafted models have made attractive enhancements in robust feature construction, and have propelled the person re-identification research. Some features based on covariance descriptor (19) were proposed which described an area of interest as a co-variance. of pixel features. It gives an easy approach to inter-wine different properties, *e.g.*, color and texture surface, of image pixel into the solitary meta-descriptors. As we know that covariance descriptor is acquired by taking average of the features inside the local area, it cures impacts of noise-clamor and spatial misalignments. And thus, it effectively handles person re-identification (20) (21) (22).

We construct a n-dimensional random vector for pixel feature depiction, $X \in \mathbb{R}^n$, where X may contain three color channel (RGB) data. Then spatial data of each pixel, including x, y values and gradient values are accumulated in X. Then more specifically X may take the representation as:

$$X = [x, y, R, G, B, |I_x|, |I_y|, \sqrt{I_x^2} + \sqrt{I_y^2}] \qquad 1.2\ (i)$$

where x. and y are horizontal and vertical positions, respectively; R, G and B are values of the three color channel; I_x and I_x are the gradients in two direction.

We then present data as a pdf. One of the ways used to estimate a pdf is histogram. However, a structure of the space framed by histograms is difficult to evaluate. Along these lines, it is difficult to design powerful similarity functions and learning algorithms for histograms. Another elementary property of pdfs are covariance matrices, and so other way to describe the signal is by utilizing its co-variance matrices as region co-variances. We have different shapes of pdf curves and surfaces delivered by different co-variance matrices. Those which are symmetric positive definite (SPD), then tend to be a connected Riemannian manifold. In this way, successful algorithms using region co-variances can be designed by examining its structure utilizing Riemannian geometry theory. But we cannot ignore the fact that region co-variance is only a partially parameterized multivariate Gaussian, which disregards the importance of mean values. If a feature vector X is contemplated with a complete parameterized multi-variate Gaussian then more viable feature descriptors can be designed as compared to region co-variances. So, building such robust feature descriptors is still an open issue in re-ID problem.

The other focused part in our work is to find ways to learn an efficient distance function so as to manage complex match and classification challenge. Most re-ID methodologies are formulated as the supervised metric learning algorithm where a projection matrix P is looked for such that the projected Mahalanobis-like distance (23) $D_M\left(x_{ik}, x_{jk}\right) = \left(x_i - x_j\right)^T M(x_i - x_j)$, where $M = P^T P$, is minimal when feature vectors x_{ik} *and* x_{jk} refer to same individual and large in the opposite case. Numerous distance evaluation methods have been put forward by considering this viewpoint (24) (12) (25) (7) (26) (27) (28). Practically, many metric learning methods given earlier (12) (6) (25) (26) demonstrate a two-step working where firstly, an algorithm called Principle Component Analysis (PCA) is used to reduce dimensions and then our matching work is analyzed on this subspace using some metric learning. But this type of two-level implementation is optimum only when the original feature dimensions are high, since otherwise for a low-dimension feature space there might be cluttering of samples from different classes after the first stage of dimension reduction itself. In LOMO (29), a subspace and metric learning method was proposed named as Cross-view Quadratic Discriminant Analysis (XQDA) which we have used for discriminant subspace and metric learning..

1.3 PROPOSED METHOD

We have considered the problem of Re-ID as the retrieval or recognition task: Given single or multiple image frames of an unknown person (query) and a gallery set that consists of a number of known or unknown people, the objective of the method used is producing a list of ranks of all the individuals in the gallery data group based on their visual similarity to the unknown person. The expectation is that the n-highest ranked matches in the gallery will provide an identity for the unknown person, thereby identifying the probe.

To tackle the various downfalls of other methods, we have proposed a framework with multi-layer representation of pixel features using Gaussian distribution. Earlier a hierarchical co-variance model has already been used effectively for image identification. In any case, important discriminative information provided by the mean data of pixels is missing in co-variance model. This issue is taken care of in this work by portraying local regions in an image by means of two level Gaussian distribution in which both the mean and co-variance characteristics are incorporated.

Particularly, we first compactly extract local patch features in a part based region model and then describe the region by parameters extracted from local patches falling under that region. As we know that a human in an image has different body parts, each part having some intra-similarity in features. So to utilize this local structure in a person image we have roughly divided the image into horizontal strips called regions. Then to create a deeper structure we have also worked on smaller consistent patches in the image with (k x k) pixel neighborhood. Appearance of each local patch is characterized by a Gaussian distribution which we refer to as *patch level or first level Gaussians.* Now we are left with multiple Gaussian distributions of patches, and then on a fixed and defined set of patches in a region we model our *region level or second level Gaussian* distribution for each region. Finally the vectors of second level Gaussian are concatenated and used as a descriptor to analyze the image. First level Gaussians are constructed on some kind of pixel features which can be color or texture data. For this we have used some kind of statistical methods to extract varied information in a pixel and then created d-dimensional pixel features. Details of this are given in chapter 4 and 6.

Consequently, we have learned a discriminative and compact model on Gaussian distributions and now we need to learn distance between shrewd Gaussian segments to measure the separation between two image sets. However, a Gaussian distribution lies on a particular Riemannian manifold, as per information geometry, where we cannot apply Euclidean functions (30) while most existing discriminant metric methods just work in Euclidean space.

So, Riemannian manifold needs to be locally flattened into a Euclidean space. This can be done by mapping it onto a tangent space favored by Riemannian metric functions. There's this Riemannian metric called the Log-Euclidean metric (31) which does this work on Symmetric Positive Definite (SPD) matrix and renders us with an effective method for projecting some point lying on the manifold onto a Euclidean tangent space utilizing the operator called matrix logarithm operator. The abilities of Log-Euclidean metric can be harnessed well if our Gaussian distribution at each level is embedded in the SPD matrix as described in the work (32). SPD matrix lies on a space which is also in fact a Riemannian manifold. So we first embed our Gaussian distribution in the SPD matrix and then apply LEM and half-vectorization approach to convert each SPD matrix logarithm of size $d \times d$ into a $d\,(\,d + 1\,)/\,2$ size vector.

For our final stage of learning a metric, we have utilized a metric proposed in (29) i.e. XQDA which learns a distinctive low dimensional subspace, and simultaneously a QDA metric, on the learned subspace. This metric formulates the problem as a Generalized Rayleigh Quotient, and obtains a solution utilizing Eigen value decomposition. This framework is proven to be highly efficient and compelling for re-ID task by investigations on four challenging databases.

1.4 MOTIVATION

The idea to use deep levels of extraction in the model originated from the understanding of distinct information stored in the local structures in a person image and the need to harness it. The people' clothing is usually in a way that differentiates the local regions like head, middle body, legs and arms. Each of these neighborhood parts are mostly isolated by contrasts in color or texture. The way these parts are spatially arranged adds up to determine the global model. However, a large set of existing descriptors (33) (34) (35) (36) (37) (19) describe the global evaluation of pixel features inside defined image divisions and loose some important information imbibed in the local structure. Conversely, the method presented in this work is portraying the global distribution utilizing the localized structure distribution of the pixel features.

Now proceeding to why we used Gaussian distribution. As it is known that mean data contains relevant information about the local parts, so it is equally important to be considered while considering co-variances in an area. And the best known way to use mean and co-variance in a single model is Gaussian distribution. Many hierarchal co-variance descriptors were proposed (38) (39), but they lack valuable mean data. As examples of parametric estimation, some works modelled these subareas of images using Gaussian mixture model (GMM). A similarity between two distributions was defined by a kernel function proposed by them, and was used on a support vector machine (SVM). In addition, Zhou *et al*. (40) used the characteristics of a GMM model constructed on each image as the visual features. Such models can be understood as calculating statistics of high level on local features and sampling them. This ideally provides the most optimum evaluation of a distribution. But we cannot overlook the fact that sampling local features from every single image in the database and approximating a GMM on such a large scale is quite impossible. Therefore, (40) a hierarchical estimation of GMM was applied, and the distribution for every single image was approximated as the amount by which it was varying or deviating from the entire training collection. But since this approximation depends on the properties of training set, this method obviously cannot always give a powerful image description. So we have instead used a Unimodal Gaussian for describing the pixel features in the local patches as they usually consists of only a small number of textures and colors.

A specific and important genre of distance metrics that show great classification capabilities for some distance learning issues is Mahalanobis metric learning. Its aim is mainly to emphasize relevant dimensions and discarding the irrelevant ones by finding a global mapping of the feature space which is at the same time linear as well. Thus, resulting in the reduction of feature vector dimensions for reduced computation in classification and analysis of the image sets. The class of Mahalanobis distance functions and the set of multivariate Gaussian have a bijection in between them usually described by corresponding co-variance matrix. Distance learning and classification problems utilizing a Mahalanobis metric learning are being the center of attraction recently. They include Large Margin Nearest Neighbor Learning (LMNN) (41), Information Theoretic Metric Learning (ITML) (24) and Logistic Discriminant Metric Learning (LDML) (42), which are at the top of chart recently. Considering the consistently developing large chunk of valuable data has raised the issue of scalability and the required level of supervision to learn any metric on a heavy range of dataset.

To meet these requirements, KISSME algorithm was learnt as a reliable metric working using coequality constraints. These constraints are interpreted as inherent inputs to the class of distance metric learning algorithms. The method which we are considering is also inspired by the same statistical inference perspective following likelihood-ratio test and is an extension of KISSME and Bayesian Metric methods. This metric does not present the problem of over-fitting and can be obtained easily with good efficiency. It was first described in the LOMO (29). It does not involve tedious iterative computation for optimizing the procedure. It just needs to compute two small sized co-variance matrices and is thus scalable to large datasets.

CHAPTER 2

RELATED WORKS

Recent efforts in this field are focused on developing discriminative features and learning distance models, or both, for robust matching. In general, more focus is given on two main areas of solution:

1. Designing a discriminative and robust appearance descriptors to describe a person's appearance.
2. Learning appropriate similarity function that maximizes the chance of getting a true correspondent.

2.1 FEATURE DESCRIPTORS

Generally, feature extraction methods can be branched as two broad sets of methods. The first set is appearance dependent and includes all those methods that try to utilize properties like color, texture and other appearance factors inherited by the image. While the other set is collection of those methods that investigate gait and movements of a person through the image frames. Their use is restricted though, when we are working on and considering varied view-point information from the camera as the gait appears to be changing from different angles. According to literature, appearance dependent methods have better discrimination capabilities and also readily available features which make them more appropriate than the other set despite being faced with pose and illumination challenges.

We know that the humans have a non-rigid body and is broken into a number of distinctive parts which gives the idea to construct part based models on them. Additionally, what is seen is that this also affects the clothing pattern (e.g., the upper and lower body parts are mostly clothed separately) and this gives distinct local information. To harness this information **part based body models** are being adapted recently. They can be further categorized roughly into three divisions:

- Fixed models, have pre-defined fixed regions describing different body parts positions.
- Adaptive models, they are subdivision models that are also pre-defined.
- Learned models, that first take a labelled training set of images and then learn model constraints (e.g., relative parts disposition).

2.1.1 LOW DIMENSIONAL FEATURE

They are formed by simple features with few dimensions and can describe a body part globally or locally as per the requirement.

Global features usually analyze the entire image as a whole set and derive a vector of fix size on image. Likely the most extensively utilized global feature is color histograms (43) (44). Given a color image of size N = W × H pixels, quantization of colors is first carried out into B bins 1, . . . , B. And then the histogram is built by counting the occurrences of a color as bin value. They may use different kinds color spaces as per the need, also evaluated by Du et al. (45). Color histograms on one hand provide decent robustness and also invariance to scale, but on the other hand are affected by change in saturation levels, illumination and color responses of the camera sensor. Probably the simplest solution to this is normalization of color data (45) achieved by dividing color channel values by summation of all channel values of that pixel. Another method was also proposed to handle these issues by Piccardi and Cheng (46) which created a histogram using top N color values and called it as Major Color Spectrum Histogram (MCSH).

Usually the peripheral pixels carry lower correspondence to the actual person information and mostly represent the background data. This issue was also handled in (2) (1) by rendering them with lesser weights than the pixels surrounding the vertical symmetric axis of person. One of the methods used was Dominant Color Descriptor (DCD) (in some methods called Representative Meta Color Model RMCM) of MPEG-7 in (47) (48), which gives the compact depiction of max illustrative hues.

Global features do not just comprise of color characteristics but also gradients, textures and repetitive patterns in the image as a whole. There are Gabor filters that detect vertical and horizontal lines in the image and Schmid filters (49) that try to capture the circular gradients. In whole they are orientation sensitive filters that extract texture and edge information on the image. They have been effectively used in other appearance descriptors as well (50) (44) (16) (5) in fusion with some color-related features. Here it must me pointed out that texture describing features are usually not distinctive enough when used alone for person re-identification problem thus must be used along with color ones.

Local features referred to an appearance characteristic over a limited area in the image (like neighborhood of a pixel). Each bounded and constrained neighborhood is then depicted as a feature vector which can be a histogram or any other suitable vector.. In the end we will be left with a collection of such local features. One of such classes is interest points like SIFT (Scale Invariant Feature Transform) (51), a famous local points descriptor. In this an interest point operator is applied to choose salient points in image that are invariant over varying scales and rotations. This work is done by computing a convolution between a 2-D Gaussian function of σ standard deviation and the image pixel value at some location co-ordinates x and y to detect out scale-extrema locations on different scales σ, and is denoted as : $L\ (x,\ y,\ \sigma) = N\ (x\ ,y\ ,\ \sigma) * I\ (x\ ,\ y)$ where $*$ is the convolution operation and N (x, y, σ) is the Gaussian referred. Then by utilizing functions such as Difference-of-Gaussians again in convolution with image finally stable key-points are detected in each scale space. Some other approaches utilized Maximally Stable Color Regions (MSCR) described in (11). It first finds some local regions showing maximal chromatic distance by performing constrained-agglomeratve-clustering. Then we depict these local clustered areas by 9-D feature vector formed by combining the details of their statistical characteristics.

Recurrent Highly-Structured Patches (RHSP), instead focuses on extracting repetitive patterns from clothes of a person. Firstly, it extracts small patches randomly which may be overlapping most often. Then those patches which are not having any gradient information, like the ones having uniform color distribution, are thrown out from consideration by calculating the patch entropy and then applying some threshold on it. Out of the residual patches, the one which are rotation invariant are kept and rest again discarded. Second, around the region where the selected patch is located a Locally Normalized Cross-Correlation function is applied to evaluate each patch recurrence. Thirdly, those patches with patches large degrees of recurrence are accumulated to maintain a final accumulation of the patches closest to the centroid. Then finally LBP histogram is used to represent these patches

Many person re-identification methods based on appearance descriptors as mentioned above used only one kind of features, either texture or color or interest points. But, as studied, combining different areas of information, that are complementary and captures different aspects of problem (like color and texture), usually tend to achieve a better performance. So such descriptors were also described in many works. In (12), color histograms in two color spaces is combined with Gabor and Schmid filter responses and extracted on strips of predefined height and position. Similarly, in (52) Color histograms in HSV space are fused with LBP hists, that capture the texture and repeated patterns, to depict partly overlapping rectangular patches sampled from image.

2.1.2 HIGH DIMENSIONAL FEATURES

Many works done on establishing robust features have also constructed high dimensional feature descriptors like Symmetry-Driven Accumulation of Local Features (SDALF) (14). It is also a part based model but also considers the symmetric element present in human body parts through the vertical axis to handle view-point variations. Then we have a method called unsupervised salience learning (52) which focuses on the rare objects (like some specific colored coats, baggage, umbrella, etc.) present in a person image that may differentiate it from other persons. It then evaluate these rare patches while matching two images. Ma et al. (16) performed some kind of conversion of local descriptors into a Fisher Vector to construct a global representation. of an image. While Cheng et al. (2) concluded the task by using color information of separate body parts and color displacements which were called as Pictorial Structures.

From the study on supervised metric learning approaches used widely in recent research areas it is inferred that they work well when fed with simple high dimensional feature that contain a lot of data irrespective of whether this data is discriminative or not. So, some less discriminative but high dimensional features were also proposed like [32, 42] where LBPs, SIFTs and compactly sampled color histograms were used.

A Covariance-of-Covariance feature (38) was proposed, in which region co-variance is approximated over local patch co-variances of pixel features. This is one of the source of motivation for the use of two level distribution model in this work. Co-variances are basically low-level statistical property of the image and thus are predicted to be stable. However, main issue is that they neglect the importance of mean data. But as we know that mean data also captures the local representation, thus we have added it to co-variance property to improve our representation so as to form a Gaussian distribution model. Earlier many works have used Gaussian distribution models such as Gaussians of Local Descriptors (GOLD) (37), Shape of Gaussians (34) and Global Gaussian (36).

However, Gaussian distributions lie on a particular Riemannian manifold as described in information geometry (30) that do not allow Euclidean operations on it. So to be able to use metrics with Euclidean functions some works (53) are focused on transforming the Gaussian matrix to an SPD matrix which can be further mapped onto a tangent space where we can apply Euclidean functions. This further mapping to a tangent space is required to be done because SPD matrices also in fact lie on a Riemannian manifold (31) (54) (55) and not on a vector space as desired. So to get the desired space to be able to work on it easily, some Riemannian metrics were proposed on the SPD manifold. One of the proposed metric was Affine-Invariant Metric (AIM) (55). It is effective in its approach but its computational cost is high because of the curvature of SPD manifold and thus is not often used in practice. Then later (31) introduced a new Riemannian metric named Log-Euclidean Metric (LEM) which allows the space of SPD matrices to be equipped with a bi-invariant metric using a Lie-group structure. This reduces the Riemannian manifold to a flat space. It evaluates only classical Euclidean computations on the domain of SPD matrix logarithm (i.e., the tangent space at identity matrix on SPD manifold), thus drastically reducing the computation time unlike AIM framework while preserving good theoretical properties.

So we have also used this approach where we first derive SPD matrices from our Gaussian model and then using the established combination of log Euclidean metric with half-vectorization of the matrix logarithm we project the manifold to a flat space as in the works (33) (37). Similar coding of Gaussian from low-level pixel features has been implemented for the task of person re-identification in (35), but they lack a multi-layer structure that we are using.

2.2 METRIC LEARNING METHODS

Besides deriving high quality features, research works are also focused on the second important task for person re-identification, i.e. metric learning (24) (12) (25) (7) (26) (27) (28) which mainly aims at maximizing the probability of a true match pair being closer intrinsically to each other than to a wrong sample.

Some of these works are mentioned here. One is these works is (5) where the capabilities of Adaboost algorithm is utilized to derive a strong classifier from number of two-class classifiers, considered weak, with each one associated to only one type of feature vector. Then another method is proposed in (50) where a linear function is derived by training a group of RankSVM method (56). This function is then able to measure the absolute difference between samples for some pair-wise relevance constraints. Further (57) proposed a Probabilistic Relative Distance Comparison (PRDC) technique to maximize the probability of a true match pair being closer intrinsically than to a wrong sample. In (16), a method called (58)Pair-wise Constrained. Component Analysis (PCCA) was used to learn a metric which can project the data points onto a low dimensional space where the distance computations between data points follow the constraints of the reduced space. Hirzer et al. (28) tried to acquire a simplified Mahalanobis metric learning by relaxing the PSD constraint on it, without compromising its performance. Li at al. (7) gave a joint model of distance metric with a locally adaptive threshold rule which he named as Locally-Adaptive Decision Functions(LADF), and used it for person verification purpose. Prosser et al. (50) defined the issue of re-identification as a ranking task, and applied RankSVM on it to learn a subspace.

Then recently LOMO (29), proposed XQDA algorithm in its work which can learn a lower dimensional subspace matrix and a metric kernel on that subspace simultaneously by following generalized Rayleigh-quotient formulation. It is closely associated and motivated by Bayesian face (59), KISSME (25), Linear Discriminant Analysis (LDA) (60), local fisher discriminant analysis (LF) (6), and CFML (13). It is described in (29) as more of an extension to Bayesian face and KISSME. The LF algorithm applies FDA together with PCA and LPP to determine a low dimensional yet discriminant subspace. The CFML also learns a comparable subspace to XQDA but altogether points to a different task. However, both LF and CFML, after deriving a subspace, simply uses the Euclidean function on it, while XQDA further approach to use an efficient metric as well.

Metric learning and similar methodologies reliably help in accelerating re-identification performance. But it must be noted that these mentioned strategies require a labelled set of training data that must be fixed in size and cannot be altered while performing the algorithm.

CHAPTER 3

BASIC FRAMEWORK

The system presented is designed to recognize any person from a noisy image given as a query. Input to the system is a blur test image acquired by a scanner or a digital camera, and its output is the person re-identified image from the stored database.

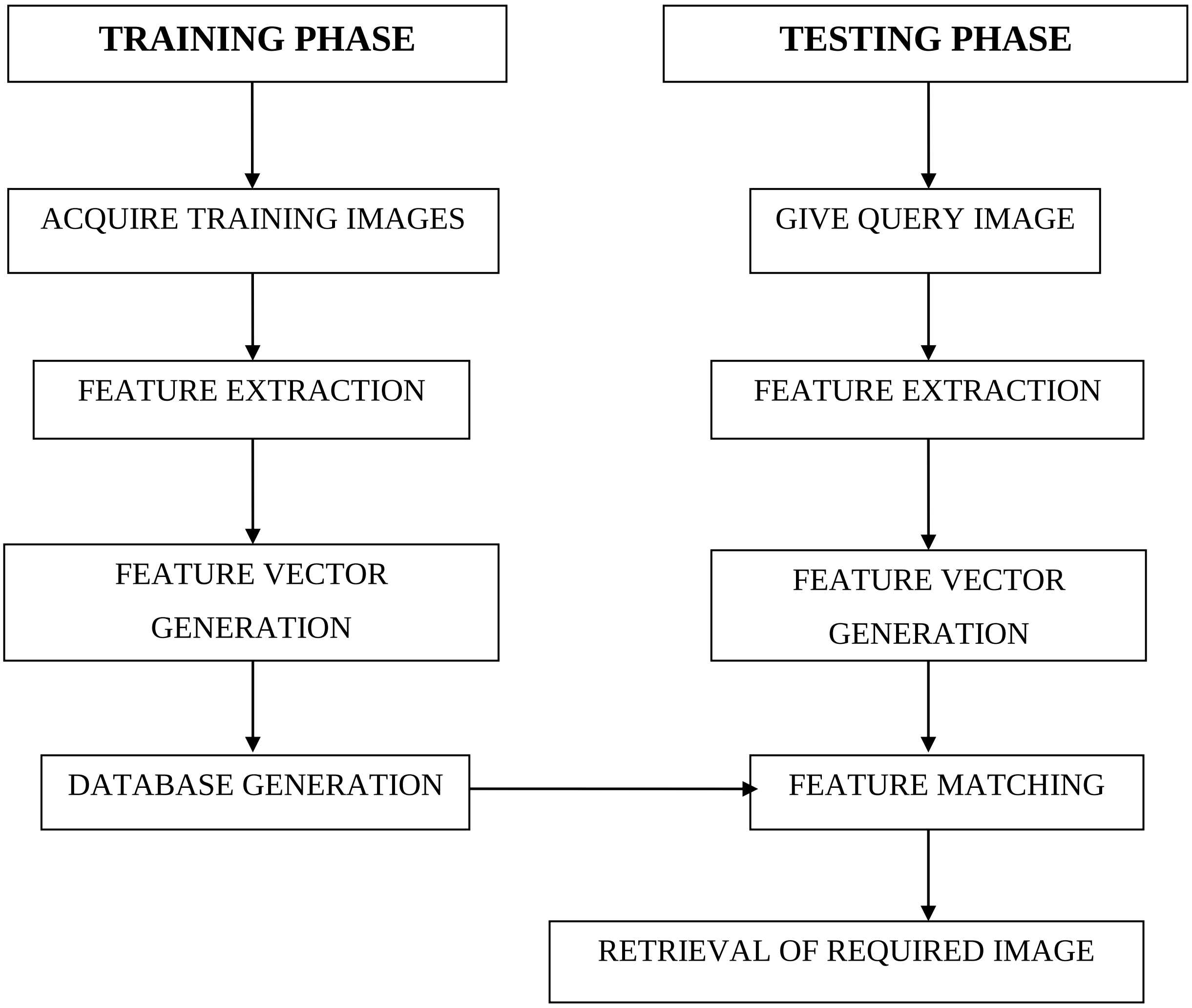


Fig. 2 Person Re Identification System Flowchart

3.1 BASIC MODULES

It is a type of image retrieval system where we try to find the correct matches of the probe image in a set of gallery images and thus retrieve correct reoccurring instances of a person from massive databases of images from multiple camera images/videos.

The system consists of the following modules:

1. Training Phase from different camera images
    a) Human Detection
    b) Image Pre-Processing
    c) Descriptor Construction
    d) Database Generation
    e) Training the Classifier/Metric Learning
2. Testing Phase for a probe image from some other camera
    a) Image Pre-processing
    b) Descriptor Construction
    c) Distance Metric Matching/Ranking
    d) Image Retrieval

3.1.1 PERSON DETECTION

If the data we have is a video sequence then first the person must be detected out of all the objects in the video frame and then further tracking is done using suitable algorithms. Picking out humans in captured images is a challenging task mainly because their body is non rigid and thus they have more than one shape models. So firstly a robust feature set is desired that can distinctly separate humans from other objects, even in cluttered scenes under difficult illumination. One of the methods which perform this task quite impressively compared to other existing methods is Histogram of Oriented-Gradient (HOG) (8) descriptors. It detects the human model by extracting gradient information at different scales of resolution applying a sobel convolution mask on the image. Then the pixels corresponding to this model are segmented from the scene in each frame using some kind of fore-ground extraction algorithms.

3.1.2 PRE-PROCESSING OPERATIONS

They are required to alter the nature of the image, which makes extraction of features easier. The point to perform pre-processing of images is to smother any undesired contortions or improve the quality of image features that are critical for further evaluation. It significantly improves the performance of the recognition system. It includes:

- *Image Resizing*

The initial form of image captured from the scanning device is usually having high dimensions and also different devices may have different resolution settings. So, to ease down our computational complexity and also to work on uniform dimensional images, we perform image resizing. This is usually done by some kind of interpolation techniques which can perform operations like zooming, rotation, shrinking and geometric corrections.

- *Image Smoothening*

When we use a scanner for image acquisition then due to sensitivity issues of the scanner some kind of noise may get added in the image. Removing this undesired noise component is an important task in this image pre-processing, because this noise may affect segmentation and pattern matching. Convolution method is used for image smoothening. In this method some transformation or statistical operation is applied on the neighborhood of the target pixel and this value then replaces the original pixel value. We may also use median filters.

So if we have a dataset of images then the images are first resized to fixed dimensions and then enhanced using filter operations before getting further processed. Most methods presume that these basic elementary steps of detection, tracking and segmentation are performed on the dataset already and then focus only on the rest of the tasks for re-ID. This work also gives importance to only the third and fourth step in the re-identification system, i.e. learning efficient descriptor and a metric.

3.1.3 DESCRIPTOR CONSTRUCTION

For person re-identification usually the non-rigid behavior of human body is taken care of by applying some kind of part based model. Then the descriptor is constructed by either estimating local or global features, or both from the silhouettes of people. This representation is then finally saved in a database to be used further in search algorithms.

These type of descriptors first divide the human body into subparts using segmentation methods or simply divide the image into regions which may be non-overlapping or overlapping based on a rough estimation of different portions of human body in the image. These regions may be fixed in size and position or adaptive to the human subparts (after being derived by an algorithm). Once we are done with this we represent these parts by global features or a. bag of local features. In this work we have used fixed part based model. The image is subdivided into seven overlapping horizontal strips of equal size which we are referring to as regions, that is approximately capturing the subparts of body, i.e. head, upper and lower torso and upper and lower legs.

3.1.4 METRIC LEARNING

Then the final step is comparing the derived features to find the most similar resemblances and so to get successful results, we require a good distance metric. Usually these metric learning methods are categorized as supervised learning and unsupervised learning, global learning and local learning, etc. In person re-ID, most of the works utilize supervised global distance metric learning techniques. They generally try to minimize the distance values between the same class objects while maximizing those from different classes. The most commonly used algorithms embrace the class of Mahalanobis distance functions, where the Euclidean distance is generalized through linear scaling and rotations of the feature space. The squared distance between the two vectors (23) $x_i\ and\ x_j$ can be given as:

$$D_M\left(x_{ik}, x_{jk}\right) = \left(x_i - x_j\right)^T M(x_i - x_j)\ , \qquad 3.1.4\ (i)$$

where M represent the learnt metric kernel.

Then some other functions, used widely to find the most similar individuals by comparing their feature vectors, are quadratic distance' sum on all points, sum of absolute differences and the correlation coefficients. But Mahalanobis distance overshadows all these by considering number of correlations among all feature vectors. But all of these simple metrics are restrictive in nature and non-flexible in a way that they consider all the features with same importance and have no intelligence to eliminate useless ones. Since some features may be more distinctive than others due to severe variations of illumination, pose and view point in the scene and thus they must be given more weight while weaker ones must be discarded.

So to overcome these shortcomings, distance metric learning techniques are being focused on by researchers like SVM, PRDC, etc. The general aim of SVM is to optimize an objective function and select the hyper-plane that separates two classes with maximum margins. While in PRDC, the objective function maximizes the probabilities of a true match pair being closer intrinsically to each other than to a wrong sample.

3.2 PROPOSED METHOD

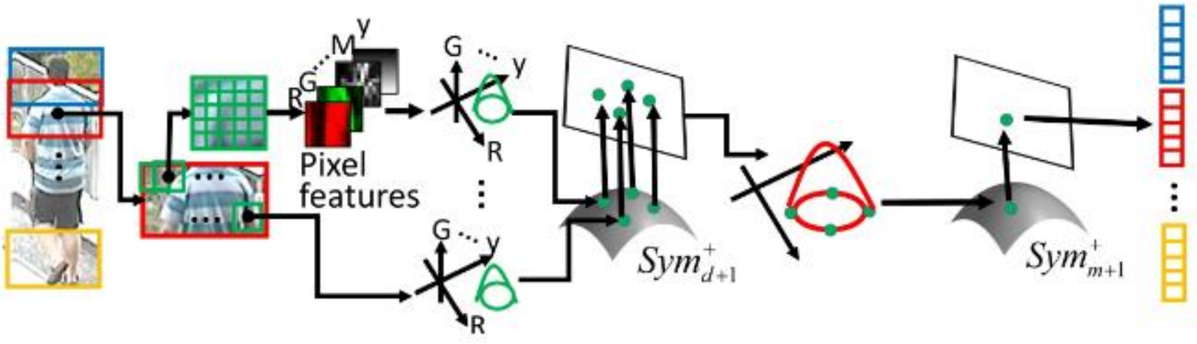
In this work, we have presented a framework for re-Id problem and in the process we have created a deep structure of Gaussian model using two layers of Gaussian evaluation for re-identifying people across challenging datasets taking motivation from (61) (29) (35) and (38). We have called this descriptor as multi-layer Gaussian descriptor.

Particularly, we first compactly extract local patch features in a part based region model and then describe the region by parameters extracted from local patches falling under that region. As we know that a human in an image has different body parts, each part having some intra-similarity in features. So to utilize this local structure in a person image we have roughly divided the image into horizontal strips called regions. Then to create a deeper structure we have also worked on smaller consistent patches in the image with (k x k) pixel neighborhood. Appearance of each local patch is characterized by a Gaussian distribution which we refer to as *patch level or first level Gaussians.* Now we are left with multiple Gaussian distributions of patches, and then on a fixed and defined set of patches in a region we model our *region level or second level Gaussian* distribution for each region. Finally the vectors of region Gaussian are concatenated and then used as a descriptor to represent the image. First level Gaussians are constructed on some kind of pixel features which can be color or texture data. For this we have used some kind of statistical methods to extract varied information in a pixel and then created d-dimensional pixel features. Details of this are given in chapter 4 and 6.

However, there is an issue that we have overlooked. A Gaussian distribution lies on a particular Riemannian manifold, as per (30)information geometry, where we cannot apply Euclidean functions. While, to estimate our second level Gaussian distribution on a region we need to evaluate two parameters of a Gaussian, i.e. the mean and co-variance matrix (mathematical operations), on a set of first level Gaussians where mathematical functions cannot be applied simply because they do not lie on a Euclidean space. So, Riemannian manifold needs to be locally flattened into a Euclidean space. This can be done by projecting it onto a tangent space favored by Riemannian metric functions. There’s this Riemannian metric called the Log-Euclidean metric (31) which does this work on Symmetric Positive Definite (SPD) matrix and thus renders us with an effective method to map a point on the manifold to a Euclidean tangent space via a matrix logarithm operator. The abilities of Log-Euclidean metric can be harnessed well if our Gaussian distribution at each level is embedded in the SPD matrix as described in the work (32). SPD matrix lies on a space which is also in fact a Riemannian manifold. So we first embed our Gaussian distribution in the SPD matrix and then apply LEM and half-vectorization approach to convert each SPD matrix logarithm of size $d \times d$ into a $d\,(d+1)/\,2$ size vector.

Consequently, we have learned a discriminative and compact model on Gaussian distributions and now we need to learn distance between shrewd Gaussian segments to measure the separation between two image sets. But again our final descriptor is a Gaussian model which lies on a particular Riemannian manifold and most existing discriminant metric methods just work in Euclidean space. So we further apply this flattening space geometry approach to transform the space of second level Gaussians. Finally our proposed descriptor is obtained by concatenating all the feature vectors, derived for separate overlapping regions in the image, to give a single vector representation of the image as depicted in Fig. 3.

Once we have constructed the feature vector representation of all the images, we have then used a discriminant distance metric to evaluate these features to find the closest matches to our probe image. We have used metric given in (29) as our discriminant metric. It learns a discriminant low dimensional subspace and a QDA metric on the derived subspace at the same time. This metric formulates the problem as a Generalized Rayleigh Quotient, and obtains a closed form solution by the generalized Eigen value decomposition.



(a) Pixel feature extraction (b) Patch Gaussian layer (c) Flatten Patch Gaussian (d) Region Gaussian layer e) Flatten Region Gaussian (f) Feature Vector

Fig. 3 Multi-layer Gaussian descriptor model

CHAPTER 4

MULTI-LEVEL GAUSSIAN DESCRIPTOR

4.1 PIXEL LEVEL FEATURES

The basic element of an image is a pixel. It has immense amount of valuable data and what to use for our work depends on the need. Sometimes we use the color information and thus depict the pixel by its color values or we can use texture data in that pixel or gradient of its value with respect to adjacent pixels and the list goes on. In this work we have used a fusion of features and have depicted each pixel i as a d – dimensional feature vector $\boldsymbol{f_i}$. This feature vector can be any such combination of features. We have performed our work on four different kinds of feature vectors where we have used the most sorted out feature, i.e. color values and color moments, texture features including Schmid filter responses and gradient values and spatial location of pixels. Their detailed explanation is as follows:

4.1.1 SPATIAL FEATURES

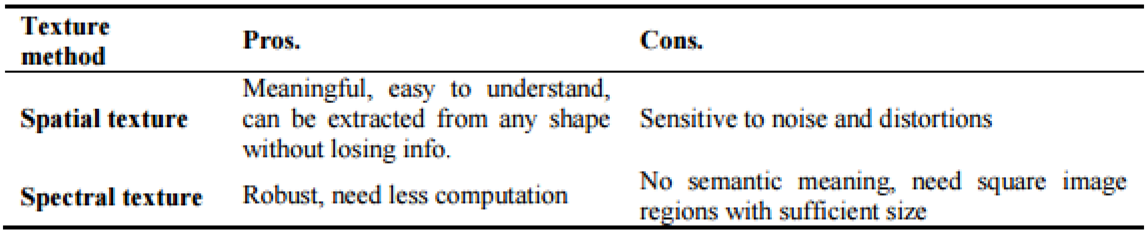
To credit the spatial information inherited by the pixels we have used pixel location as one of the element in pixel representation. As we know that the human body parts are symmetrical across the vertical axis and always follow a fixed representation in vertical direction, i.e. first comes the head then torso and then legs, and that is universal and do not change with any view point variations or little misalignments, while this cannot be said true for the body model in horizontal direction. So it is more trustworthy and distinctive to use the spatial location of a pixel in vertical direction (35) as our spatial feature rather than unreliable horizontal spatial positioning. For each pixel i this is defines as y_i ,considered in reference to the highest point of the image.

4.1.2 TEXTURE DATA

It is believed generally that our capabilities of recognizing the minute details of the scene and its understanding comes from the ability to process texture information and thus it is an important characteristic of a pixel. It can be broadly understood as a measure of variation of intensity values through the image surface. As already discussed in the related works section of this text that texture as a feature is not self-sufficient in depicting the person images distinctly but can separate out textured images from non-textured ones. But when used along with some color description can prove to be effective.

When texture features are evaluated by deriving a statistical distribution on intensity values of the pixels in an image then methods are called *statistical methods*. Some of these methods are namely, Fourier power-spectra., co-occurrence matrices, shift in-variant principal component analysis (SPCA), Tamura features, Wold decomposition, Markov random field, and also include some multi-resolution filtering techniques like Gabor and. wavelet trans-forms. They are also extracted either directly on the local surface level of the image or in the orientation level also called frequency domain of image. Then depending on the domain of evaluation, they are broadly classified into spatial domain texture methods and spectral domain texture methods. They both have their own advantages and disadvantages as summarized in Table 1.

Table 1. Comparison of different textural feature extraction methods



Texture method	Pros.	Cons.
Spatial texture	Meaningful, easy to understand, can be extracted from any shape without losing info.	Sensitive to noise and distortions
Spectral texture	Robust, need less computation	No semantic meaning, need square image regions with sufficient size

4.1.2.1 GRADIENT INFORMATION

The gradient information can be understood as the variations in the intensity values of the adjacent pixel across different directions and their derivatives. Let I defines a image region and $r = (x, y)$ gives the position vector of a point in I. Then the image gradient ($\partial I/\partial x$, $\partial I/\partial y$) at each pixel can be defined in terms of the magnitude $m = (\frac{\partial I}{\partial x})^2 + (\frac{\partial I}{\partial y})^2$ and the orientation angle $\theta = arctan (\partial I \,/\, \partial x, \partial I \,/\, \partial y)$ calculated from *x* and *y* derivatives I_x, I_y of intensity *I*. The orientation θ is quantized into D orientation bins (61), we have used four, by voting weights to the nearest bins, and is described as a sparse vector $f(\in \mathbb{R}^D)$, and then called the gradient orientation vector (in short, GO) now defined as : $O_{\theta \,\epsilon\, \{0^{\circ}, 90^{\circ}, 180^{\circ}, 270^{\circ}\}}$. These weight values are estimated linearly by utilizing the distance approximations from the quantized orientations similar to the way we calculated GO vector as in (62). We have also demarcated the high gradient value points as edges by multiplying the gradient magnitude $M = \sqrt{(I_x^2 + I_y^2)}$ to the quantized orientation O_θ and thus obtained the oriented gradient magnitude: $M_\theta = MO_\theta$.

4.1.2.2 SCHMID FILTER

Schmid proposed texture feature filter banks which have rotational invariance property and can be estimated by convolving isotropic "Gabor-like" filters (49) with the image. It is a grey value descriptor dl computed for each image pixel location. Each texture channel is computed by convolving image points with the filter values and luminance channel. These filters combine different frequencies and scales together as follows:

$$F(r, \sigma, \tau) = \frac{1}{Z}\cos(\frac{2\pi\tau r}{\sigma})\exp(-\frac{r^2}{2\sigma^2}) \qquad 4.1.2.2\ (i)$$

Here, r depicts the radius, Z depicts the number of cycles of the harmonic function within the Gaussian envelope of the filters in context of Gabor filters and is called normalizing constant. σ represents the scale of the filter.

For our experiments we have used same settings as in (5), we have used 13 filters with scales σ between 2 and 10 and τ between 1 and 4. For smaller scales only small τ are used to avoid high frequency responses. In this work, several filters are generated by taking different banks parameters (σ, τ) pairs. They are taken by ranges (2,1), (4,1), (4,2), (6,1), (6,2), (6,3), (8,1), (8,2), (8,3), (10,1), (10,2), (10,3) and (10,4), 13 filters in all and the responses of the image from Schmid filter bank are shown in Fig. 4. We are using these filters to create feature invariance to pose and view-point, while originally their creator designed them to handle rotational variations. A comparison of Schmid filters with rotational in-variant combinations of derivatives has shown that the Schmid filters outcasts them in its performance.

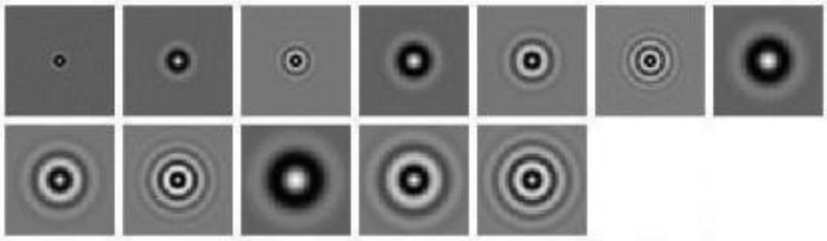


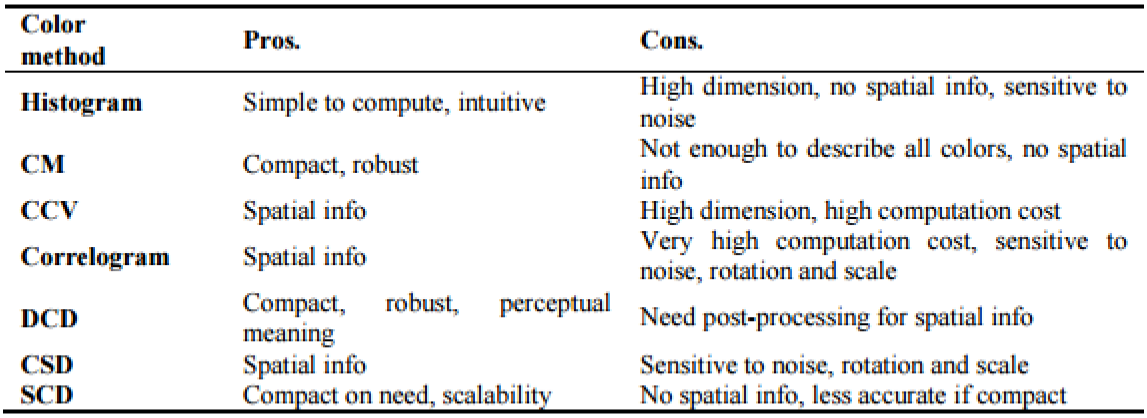
Fig. 4 Rotationally Symmetric Schmid filters

4.1.3 COLOR FEATURE

Color is basically the way we humans understand and depict a range of wavelengths, approximately 300 to 830 nm, in the electromagnetic spectrum, Color features are some property of these wavelengths that can describe an image pixel. Some already proposed and extensively used properties are color histogram, color moments (CM), color coherence vector (CCV) and color

Correlograms. Same color value can also be depicted differently just based on different color evaluation characteristics of the system and these different representations are called color spaces. Some widely used ones are RGB, LUV, HSV and HMMD. In Table 2 we enlist some of the color based features used in research works along with their merits and demerits.

Table 2. Summarized Comparison of Several Color Descriptors



Color method	Pros.	Cons.
Histogram	Simple to compute, intuitive	High dimension, no spatial info, sensitive to noise
CM	Compact, robust	Not enough to describe all colors, no spatial info
CCV	Spatial info	High dimension, high computation cost
Correlogram	Spatial info	Very high computation cost, sensitive to noise, rotation and scale
DCD	Compact, robust, perceptual meaning	Need post-processing for spatial info
CSD	Spatial info	Sensitive to noise, rotation and scale
SCD	Compact on need, scalability	No spatial info, less accurate if compact

4.1.3.1 COLOR SPACES

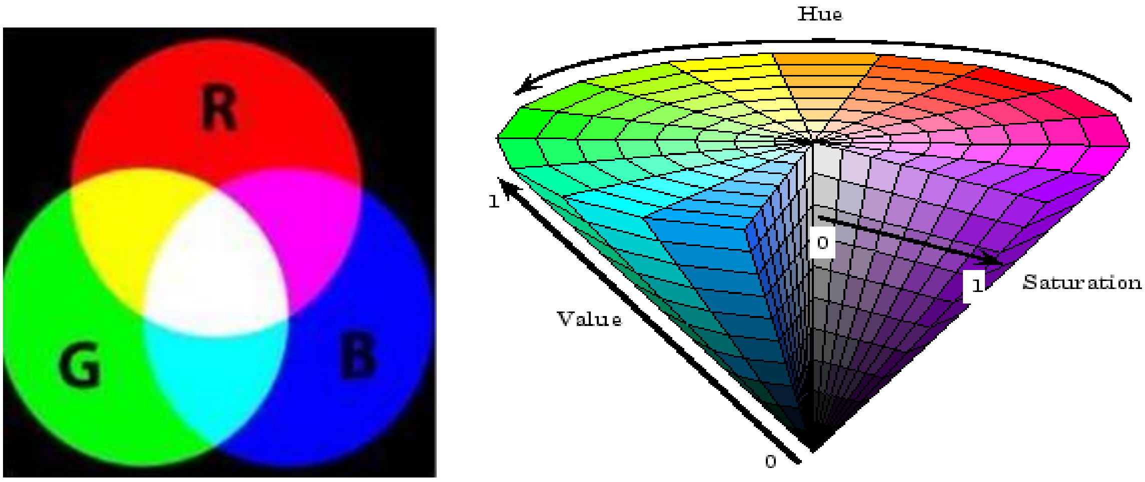
The choice of the color space sometimes can largely influence the results of processing and thus it is important to select it wisely. Various color spaces describe the same color information in different ways such that it makes certain computations easier. They left us with a distinct way to identify colors.

Some common color spaces are: RGB, CMYK, YUV, YCbCr, LAB, LUV and HSV.

In the proposed method we have utilized the representation of color information on two different color spaces, namely normalized RGB and HSV.

- RGB - RGB stands for red, green and blue color components. As stated by RGB model, each color image is a combination of intensity values of three different images, a Red, Green, and Blue image as shown in Fig. 5(a). Thus a color image matrix has three dimensions, each corresponding to one of the three colors, while a gray scale image is defined by only one dimension matrix. It can also be described as addition of three different color matrices.

- HSV - The HSV color space (Hue, Saturation. and Value) describe color information in a way that is closer to how humans perceive them than what RGB values give. It is similar to selecting colors on a color wheel or palette. As hue value varies from 0 to 1.0, relatively the colors vary from red to yellow to green, cyan, blue, magenta, and then back to red, i.e. in a circular base. As saturation varies from 0 to 1.0, the corresponding colors value or hues successively change from an unsaturated (gray shades) to completely saturated form of that color (no white component). As brightness parameter varies from 0 to 1.0, the corresponding colors become progressively brighter. Fig. 5(b) illustrates the HSV color space.



(a) RGB (b) HSV

Fig. 5 Illustration of RGB and HSV color spaces

Hue values linearly follow a high to low transition. When original image is compared with the hue plane image it clearly shows that shades of deep blue correspond to the highest possible values, and shades of deep red represent the lowest ones possible. **Saturation** can be defined as how pure a color is. The colors having higher saturation content have the highest values in the saturation plane image. **Value** roughly describes the brightness levels.

4.1.3.2 COLOR MOMENTS

As we know that moments are statistical parameters mostly used to describe a probabilistic distribution, in the same way moments on color spaces are measures that portray color distributions in an image. They have ability to distinctly depict an image and thus can give an estimation to measure similarity between image pairs. The premises for extracting moments on color spaces come from the presumption that color in an image can be depicted as a probabilistic distribution and as these moments can uniquely describe the probability distribution similarly moments on color space can also be used as discriminative features to describe an image.

We usually utilize only the first three color moments as our feature due to the fact that major distribution information is inherited by the lower order moments. So if a color space used has three dimensions like RGB space then total moments computed are 9, since it is evaluated on each channel separately. They are inclusive of both shape and color data in the pixels thus give a illumination invariant as well as scale and rotation invariant feature representation but they are weak at handling any occlusion.

If we represent the j^{th} image pixel at the i^{th} color channel as P_{ij} and number of image pixels being considered as N, then the three color moments are derived as:

- **Mean**: It is defined as the average color value of the image pixels and denoted as E_i.

$$E_i = \sum_{j=1}^{N} \frac{1}{N} P_{ij} \qquad 4.1.3.\ (i)$$

- **Standard Deviation**: The standard deviation is the square root of the variance of the distribution :

$$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij} - E_i\right)^2\right)} \qquad 4.1.3.\ (ii)$$

where E_i is the mean value, or first color moment, for the ith color channel of the image.

- **Skewness**: Skewness is defined as a measure of the degree of asymmetry in the color distribution and thus it provides some insight on the shape of the color distribution.

$$S_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N}(P_{ij} - E_i)^3\right)} \qquad 4.1.3\ (iii)$$

4.2 PATCH LEVEL GAUSSIANS

Let I be a three-dimensional color image i.e. each pixel has three-dimensional representation and F be a transformed image representation of $W \times H \times d$ dimensions where d denotes the new dimensions of each pixel feature. Then this is learned from the image I by some transformation function, such that

$$F(x, y) = \phi(\mathrm{I}, \mathrm{x}, \mathrm{y}), \qquad 4.2\ (i)$$

F can consists of any combination of feature values like intensity, color, gradients, filter responses, etc. Using this approach we have used a combination of color moments, Schmid filter responses, gradients and pixel location as our pixel feature vector. One of them is represented as:

$$\boldsymbol{f_i} = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, H, S, V]^T\ , \qquad 4.2\ (ii)$$

where y is the coordinate in the vertical direction, which can be used to keep the spatial orientation of the body, next 3 dimensions represent the gradient values and last three dimensions give the HSV space values. It is an 8-dimesion vector. Similarly we have defined a13-dimensional pixel feature using Schmid filter responses and a 7-dimension vector with color moments as it elements as:

$$\boldsymbol{f_i} = [y\ mean(R) std(R) mean(G) std(G) mean(B) std(B)] \qquad 4.2\ (iii)$$

So after we extract the pixel features inside a patch, we then summarize them via the most classical parametric distribution which has mean and covariance as parameters: Gaussian distribution.

Let F = { f_1 : : : f_D } be the set of d-dimensional local features and assume that they are independent and identically distributed, then we have demarcated the local neighborhood of (k x k) pixels as a patch and then, considering number of such overlapping patches with some fixed patch interval, for every patch 'S', the Gaussian model $\mathcal{N}(f\ ;\ \mu_s, \Sigma_s)$ is defined as,

$$\mathcal{N}(f\ ;\ \mu_s, \Sigma_s) = \frac{1}{(2\pi)^{d/2}|\Sigma_s|}\left[\exp(-\frac{1}{2}(f - \mu_s)^T \Sigma_s^{-1}(f - \mu_s))\right], \qquad 4.2\ (iv)$$

as also described in (61) where | · | is the determinant of a matrix, μ_s is the mean vector and Σ_s is the co-variance matrix of the patch s. $f, \mu_s \in \mathbb{R}^d\ and\ \Sigma_s \in \mathbb{S}_d^+$ and $\mathbb{S}_d^+$ is the space of real symmetric positive semi-definite matrices. The mean and co-variance matrix are estimated as follows:

$$\mu_s = \frac{1}{N}\sum_{i \epsilon \mathcal{L}_s} f_i \qquad 4.2\ (v)$$

$$\Sigma_s = \frac{1}{N-1}\sum_{i \epsilon \mathcal{L}_s}(f_i - \mu_s)(f_i - \mu_s)^T \qquad 4.2\ (vi)$$

Where $\mathcal{L}_s$ is the area of the sampled patch 'S' and N denotes the number of pixels in $\mathcal{L}_s$. The approximated co-variance matrix imbibes the information on variances of the features and their correlations. When combined with mean, it improves the overall representation of feature F.

This covariance matrix is $d \times d$ dimensional and is evaluated on the patch pixel points. The diagonal entries of the covariance matrix represent the variance of each feature, and the non-diagonal entries are their respective correlations. Covariance computation serves a dual purpose as on one hand it depicts the variance of feature values in a local area and on the other hand filters out the noisy samples in the patch due to the applied average filter used in its computation. Due to the symmetry observed, Co-variance matrix have only $(d^2 + d)/2$ distinct values. Referring to a patch S, its co-variance Σ_s have no information about the order and the number of points. This points to some level of invariance to scale and rotation for the related patch. But if we use some kind of orientation data, like the gradient with respect to x and. y, in the feature point representation then we can no longer see its rotation invariance property. The compactly derived mean vectors and covariance matrices can be effectively computed from integral images (10). Integral image used is constructed for the entire image and not for local regions separately, as the areas considered are overlapping in nature.

Integral images as described in (10) are intermediate image representations utilized for the fast computation of region sums. Each pixel in a integral image is the sum of all the pixels inside a. rectangle whose boundary is defined by the upper left corner of the image and the pixel at interest point. For an image I, its integral image is depicted as:

$$Integral\,Image(x' \,,\, y') \;=\; \textstyle\sum_{x \,\leq\, x', y \,\leq\, y'} I(\,x \,,\; y\,) \qquad 4.2\ (vii)$$

It provides a fast computation method where sum of any rectangular region can be computed in a constant time.

Now our next step is to compute a Gaussian distribution on a predefined region. So, we need to evaluate two parameters of a Gaussian, i.e. the mean and co-variance matrix (mathematical operations), on a set of first level Gaussians. But from the information geometry we know that Gaussian distribution lie on Riemannian manifold where Euclidean functions cannot be applied simply. So, this Riemannian manifold need to be flattened into a Euclidean space (54) (55) (31). This can be done by projecting it onto a tangent space favoured by Riemannian metric functions. So the Riemannian metric being used here is Log-Euclidean metric (31) which does this work on Symmetric Positive Definite(SPD) matrix and thus renders us with an effective method to map a point on the manifold to a Euclidean tangent space via a matrix logarithm operator. This is explained in detail in the next subsection. So it can be inferred from this that the abilities of Log-Euclidean metric can be harnessed well if our Gaussian distribution at each level is embedded in the SPD matrix, which is also considered as a Riemannian manifold, as described in the work (32).

4.3 DEALING WITH RIEMANNIAN SPACE

A manifold is a topological space that is locally similar to a Euclidean space. Each point on the manifold has a neighborhood for which there exists a homeomorphism (one-to-one, onto, and. continuous mapping in both directions), mapping the neighborhood to $\mathbb{R}^m$.

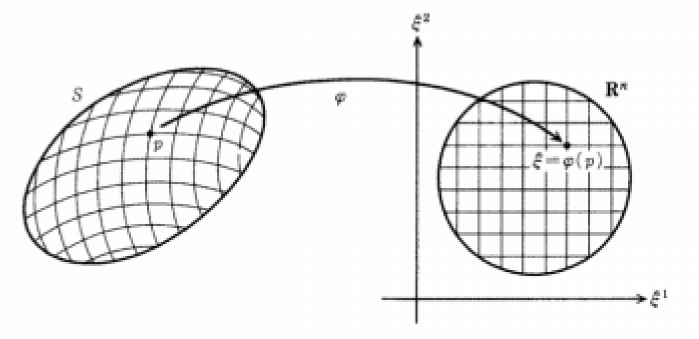


Fig. 6 A co-ordinate system ξ for a manifold S. It shows a one-to-one mapping from S to $\mathbb{R}^n$

For differentiable manifolds, described as in (10), we can define the derivatives of the curves on the manifold. The derivatives which are derived at a point X on the manifold $\mathcal{M}$ lie in a vector space $T_X\mathcal{M}$., which is the tangent space at that point.

A Riemannian manifold is a differentiable manifold $\mathcal{M}$ endowed with a Riemannian metric. In this every tangent space have inherit inner product $<,>_{X \in M}$, which transcends gradually and smoothly between points. A norm in this tangent space is denoted as:

$$\| \ y \ \|_X^2 = < y \, . \, y >_X . \qquad 4.3 \text{ (i)}$$

The minimum length curve joining a pair of given points on the manifold is called the geodesic, and the distance between a pair of points on the manifold denoted as, $d \ (\ X \ , \ Y \ )$, is delivered by the length of this curve. Let $y \in T_X\mathcal{M}$ and $X \in \mathcal{M}$. From X, we existentially have a unique curve or geodesic with its start point being on the tangent vector$y$. Then the vector y can be mapped to the end point of this geodesic by applying the exponential map exp_X : $T_X \ \mathcal{M} \leftrightarrow \mathcal{M}$, and the distance of the geodesic can be defined as:

$$d \left(X \, , \, exp_X(\, y \,) \right) = \| \ y \ \|_X \qquad 4.3 \text{ (ii)}$$

Generally this map is onto, but in the neighborhood of a point $X \ \in \ \mathcal{M}$ it is only one-to-one. Therefore, the inverse mapping $\log_X : \mathcal{M} \longleftrightarrow \mathrm{T_X}\mathcal{M}$ is distinctly established just around a small neighborhood of this point X. If for any $\mathrm{Y} \in \mathcal{M}$, there exists many $y \in T_X\mathcal{M}$ such that $Y = exp_X(y)$, then $log_X(Y)$ is produced by the tangent vector with smallest norm. This is well depicted in Fig. 7. At this point it must be mentioned that both the operators are point-dependent clearly denoted by use of subscript.

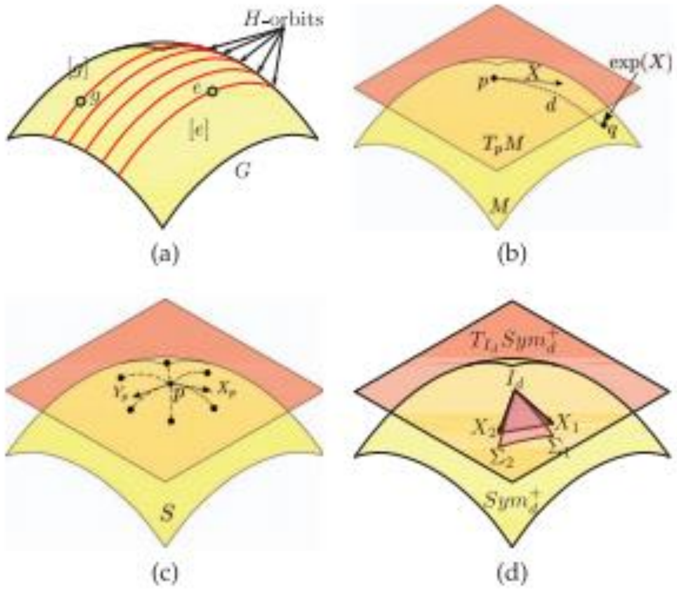


Fig. 7 Mapping in Riemannian Space: (a) Homogeneous Spaces (b) Exponential Map
(c) Gaussian curvature of the 2-D surface S at p
(d) Approximating the true geodesic distance

4.3.1 SPD MATRIX MANIFOLD

As it is well described in (54) (55) (31), that the space of $d \times d$ SPD matrices also lie on a particular Riemannian manifold when applied with suitable Riemannian metric and so is called as SPD manifold S_d^+. The SPD manifold, showing much resemblance to a Euclidean space locally, is a topological space with globally defined differential structures, which creates a possibility that curves' derivatives can be defined on the manifold by using a logarithm map:

$$log_P: S_d^+ \rightarrow T_P S_d^+ (X \in S_d^+) \qquad \text{i.e. equivalent to } log_X: \mathcal{M} \leftrightarrow \mathrm{T_X}\mathcal{M} , \qquad 4.3.1\text{ (i)}$$

These derivatives at the point P on the manifold lie on a tangent space $T_P S_d^+$, which has an inner product $<,>_P$.

Now we need to perform the embedding of Gaussian distribution on a SPD matrix. This was well described by (53) (63) in their work where they transformed a Gaussian model to an SPD matrix using the theory of information geometry (30). In the manner similar to the work (32), we have also embedded our d-dimensional multi-variate Gaussian distribution into another $d + 1$ dimensional SPD matrices space which is also a Riemannian manifold denoted by SPD_{d+1}^+ .

Let $\mathcal{N}(0, I)$ is the definition of a d-dimensional Gaussian distribution with the mean vector 0 and co-variance matrix I which is an identity matrix and $|\cdot|$ denotes the matrix determinant. Now if a random vector x follows this Gaussian distribution $\mathcal{N}(0, I)$, then its afine transformation denoted as $Qx + \mu$ also follows a Gaussian model $\mathcal{N}\ (\mu, \Sigma)$, where Σ is a covariance matrix which can be decomposed as $\Sigma = Q^T Q,\ for\ |Q| > 0$, and vice versa. Following this analysis (32), the Gaussian $\mathcal{N}(\mu, \Sigma)$ can then be characterized by a affine transformation(μ, Q). Now let us denote τ_1 as the mapping from affine group -

$$\mathrm{AFF}_d^+ = \{(\mu, Q) | \mu \epsilon \mathbb{R}^d, Q \epsilon\ \mathbb{R}^{d \times d},\ for\ |Q| > 0\} \qquad 4.3.1\ (ii)$$

to the special general linear group

$$S\mathcal{L}_{d+1} = \{A | A\ \in\ R^{(d+1)\times(d+1)}, |A|\ >\ 0\}, \qquad 4.3.1\ (iii)$$

and τ_2 denote the mapping from $S\mathcal{L}_{d+1}$ to SPD matrix space, then represented as,

$$SPD_{d+1}^+\ = \{P | P\ \in\ \mathbb{R}^{(d+1)\times(d+1)}, P\ =\ P^T\ , |P|\ >\ 0\}, \text{ i.e.,} \qquad 4.3.1\ (iv)$$

$$\tau_1 : \mathrm{AFF}_d^+ \rightarrow\ S\mathcal{L}_{d+1} \qquad \tau_2\ :\ S\mathcal{L}_{d+1}\ \rightarrow\ SPD_{d+1}^+ \qquad 4.3.1\ (v)$$

$$(\mu, \Sigma)\ \rightarrow\ C_Q \begin{bmatrix} Q & \mu \\ 0^T & 1 \end{bmatrix}\ , \qquad S\ \rightarrow\ SS^T \qquad 4.3.1\ (vi)$$

where $C_Q\ =\ |Q|^{-1/(k+1)}$. Through these two mappings, a d-dimensional patch Gaussian $\mathcal{N}(\mu_s, \Sigma_s)$ can be embedded into a d+1 dimensional SPD space denoted as SPD_{d+1}^+ and thus is uniquely represented by a $(d+\ 1)\ \times\ (d\ +\ 1)$ SPD matrix P_s ; that is,

$$\mathcal{N}(f\ ;\ \mu_s, \Sigma_s) \sim P_s = |\Sigma_s|^{-1/(d+1)} \begin{bmatrix} \Sigma_s + \mu_s \mu_s^T & \mu_s \\ \mu_s^T & 1 \end{bmatrix} \qquad 4.3.1\ (vii)$$

But a problem occurs when there are not enough pixel points within a patch and consequently the matrix representation of a patch becomes singular. So this problem is handled by avoiding it as described in (61), and thus we add the identity matrix $\mathrm{I_d}$ to Σ_s with a small positive multiplied constant value, ϵ_s, as:

$$\Sigma_\mathrm{s} \leftarrow \Sigma_\mathrm{s} + \epsilon_\mathrm{s} \mathrm{I_d}. \qquad 4.3.1(viii)$$

After this process is completed we are left with patch level Gaussian on the SPD manifold P_s which is then required to be mapped onto a tangent space via a matrix logarithm as described in the next subsection.

4.3.2 TANGENT SPACE MAPPING

As described in (54) (31) (55), that SPD matrices do not lie on a vector space but instead on a specific Riemannian manifold. Therefore, to work with metrics that operate on Euclidean structures, we need to find a suitable mapping method that can transform each local neighborhood of a point on manifold to an open set in a Euclidean space, thus projecting the data points from the manifold to a Euclidean space. Firstly we need to choose a suitable tangent space $T_X M$ on which our data is to be mapped ($X \in M$). The exponential map (exp_X) and logarithmic map (log_X) can then be utilized to define appropriate coordinates values on that space.

So now we will try to understand the Log-Euclidean metric. As observed in (31) and explained in (33) the simple matrix exponential (exp) is a diffeomorphism from the Euclidean space of symmetric matrices ($T_P S_d^+$) to the space of S_d^+. The important point is that a matrix logarithm of $P \in S_d^+$ is unique, well defined and is a symmetric matrix $u = log(P) \in T_P S_d^+$ whereas the matrix exponential $P = exp(u)$ of any symmetric matrix $u \in T_P S_d^+$ gives a matrix $\in S_d^+$. *The Log-Euclidean framework employs the simple matrix logarithm as a mapping, resulting in a space of* S_d^+ that is isomorphic (the algebraic structure of the vector space is conserved), diffeomorphic and isometric (distances are conserved) to the associated Euclidean space of symmetric matrices. The matrix logarithm can be viewed as the logarithm map with base point set at the identity matrix I_d

4.3.2.1 LOG-EUCLIDEAN METRIC

By name we can understand that it points to a Euclidean metrics in the logarithmic domain. In the Log-Euclidean framework, the logarithmic multiplication ⊙ and the scalar logarithmic multiplication ⊗ are defined such that S_d^+ has a linear space structure.

The exponential map associated to the Riemannian metric (10) is a global diffeomorphism (one-to-one, onto, and with continuously differentiable mapping in both directions).

$$exp_P(y) = P^{\frac{1}{2}} \exp(P^{-\frac{1}{2}} y P^{-\frac{1}{2}}) P^{\frac{1}{2}} \qquad 4.3.2.1 \text{ (i)}$$

Therefore, at all the points on the manifold on which we have embedded the patch Gaussians,i.e, S_{d+1}^+, the logarithm is uniquely derived and then the projected tangent vector of SPD matrix P_s is given by :

$$y = log_P(P_s) \triangleq P^{\frac{1}{2}} \log(P^{-\frac{1}{2}} P_s P^{-\frac{1}{2}}) P^{\frac{1}{2}} \qquad 4.3.2.1 \text{ (ii)}$$

The exp and log operators are the ordinary matrix exponential and logarithm operators. They are different than exp_X and log_X, which denote the point dependent manifold specific operators, depending on points $P \epsilon S_{d+1}^{+}$ to which the hyperplane is tangent.

As we know, for symmetric matrices, the ordinary matrix exponential and logarithm operators can be calculated easily, so we apply simple matrix logarithm as a mapping to give us the projected vector on an Euclidean Space. At some specified tangency matrix P, this space can be seen tangent to the Riemannian manifold. The tangent space of S_{d+1}^{+} is the space of $(d+1) \times (d+1)$ symmetric matrices, and both the manifold and the tangent spaces are $m = \frac{(d^2+3d)}{2} + 1$ dimensional. The matrix logarithm operators on our embedded SPD matrix P_s can be derived by Eigen-value decomposition of a symmetric matrix ($P_s = UDU^T$) and is defined as,

$$\log(P_s) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} (P_s - \mathrm{I})^k = U \log(D)\, U^T \qquad 4.3.2.1 \text{ (iii)}$$

Similarly, the matrix exponential series is defined as,

$$\exp(P_s) = \sum_{k=0}^{\infty} \frac{{P_s}^{\mathrm{k}}}{k!} = Uexp(D)U^T \qquad 4.3.2.1 \text{ (iv)}$$

where $\exp(D)$ denotes the diagonal matrix of the eigen-value exponentials. The exponential operator defined for all matrices, while the logarithmic operator is defined only for symmetric matrices with non-negative Eigen-values, S_{d+1}^{+} .

In the next step, some orthonormal co-ordinates are extracted from the projected vector so that the points in the tangent space can be presented with a minimal representation. Now the upper triangular or lower triangular part of the extracted matrix consists of $d(d+1)/2$ independent coefficients in the tangent space for the symmetric matrix of dimensions $d \times d$ as the tangent space itself is the space of symmetric matrices and thus the off-diagonal points are counted twice during norm estimation.

The orthonormal coordinate system is defined at point P on the tangent space via utilizing vector operation. For a tangent vector y in the tangent space at point P, the orthonormal co-ordinates can be derived by the vector operation as follows:

$$g_s = vec_P(y) = vec_I(P^{-\frac{1}{2}} y P^{-\frac{1}{2}}) \qquad 4.3.2.1 \text{ (v)}$$

where I denotes a identity matrix, and the vector operator at identity of a symmetric matrix Y is derived as:

$$vec_I(Y) = \left[y_{1,1} , \sqrt{2}y_{1,2} , \dots , \sqrt{2}y_{1,d+1} , y_{2,2} , \sqrt{2}y_{2,3} , \dots , y_{d+1,d+1} \right]^T \qquad 4.3.2.1 \text{ (vi)}$$

This must be pointed out here that the orthonormal coordinates of y acquired by this vector operator $vec_P(y)$ lie in a Euclidean space in $\mathbb{R}^m$ while tangent vector y itself is a symmetric matrix is. So it is shown how this vector operator maps the Riemannian metric (64) defined on the tangent space onto a canonical metric validated in $\mathbb{R}^m$

$$< y, y >_P = \| vec_P(y) \|_2^2 \qquad 4.3.2.1 \text{ (vii)}$$

Then by Substituting y from Eq. 4.3.2.1 (iii) in Eq. 4.3.2.1 (v), the projection vector of P_s on a hyperplane, tangent to P, reduces to

$$g_s = vec_I(\log(P^{-\frac{1}{2}}P_S P^{-\frac{1}{2}})) \qquad 4.3.2.1 \text{ (viii)}$$

So the matrix of patch Gaussian ***Ps*** becomes $m = (d2 + 3d)/2 + 1$ dimensional vector g_s. The studies show that at the point of projection P, the neighborhood relation between the data points remain same on the manifold. Therefore the best choice for P, to ease down the computations, is the point of identity matrix, which simply means to apply the vec_I operator to the standard matrix logarithm. This also leads to a formulation of a generalized descriptor as we don't need to optimize the projection points for each data under consideration, specifically.

4.4 REGION LEVEL GAUSSIAN

As we know that a human in an image has different body parts, each part having some intra-similarity in features. So to utilize this local structure in a person image we have roughly divided the image into horizontal strips called regions. Now after we are done with estimating local Gaussian patches we are left with multiple Gaussian distributions of these patches, and now on a set of patches falling under these regions we model our *region level or second level Gaussian* distribution for each region. The patch Gaussians are already flattened and thus the mean and covariance information of a set of these Gaussian distributions can be easily determined to describe the second level Gaussian. We usually don't order these patches while considering inside a region, since pose changes in the person images may change these local patch positions.

Here we have also considered the issue that some patches represent the background areas which differ from one image to other and also that mostly the person information is centrally placed in the image and thus to reduce the effects of these background data we give more importance to the center aligned patches as in (61)by adding weight parameter to every patch similar to as it was done in weighted color histograms. In this work we have defined these weights as:

$$w_s = \ exp(-(x_s - x_c)^2/2\sigma^2) \qquad 4.4\ (i)$$

where $x_c = W/2,\ \sigma = W/4$. Here x_s refer to the x coordinate of the center pixel of patch 'S' and W is width of the image. Then the new weighted mean vector and co-variance matrix can be defined as

$$\mu^{\mathcal{G}} = \frac{1}{\sum_{s\,\epsilon\,\mathcal{G}} w_s} \sum_{s\,\epsilon\,\mathcal{G}} w_s g_s, \qquad 4.4\ (i)$$

$$\Sigma^{\mathcal{G}} = \frac{1}{\sum_{s\,\epsilon\,\mathcal{G}} w_s} \sum_{s\,\epsilon\,\mathcal{G}} w_s (g_s - \mu^{\mathcal{G}})(g_s - \mu^{\mathcal{G}})^T, \qquad 4.4\ (ii)$$

where $\mathcal{G}$ represents the overlapping regions. Then incorporating both these values, i.e. the mean vector and co-variance matrix, we define the second level Gaussian distribution as $\mathcal{N}(g; \mu^{\mathcal{G}}, \Sigma^{\mathcal{G}})$.

Consequently, we have learned a discriminative and compact model on Gaussian distributions and now we need to learn distance between shrewd Gaussian segments to measure the separation between two image sets. However, region Gaussian distribution also lies on a Riemannian manifold, as per information geometry (30), where we cannot apply Euclidean functions while most existing discriminant metric methods just work in Euclidean space. So, we further apply flattening and half-vectorization on these second level Gaussians to map them to a Euclidean space. Firstly the m-dimensional region distribution (61) is embedded into a m+1 dimensions SPD matrix denoted as Q, same as explained in the previous section: $\mathcal{N}(g; \mu^{\mathcal{G}}, \Sigma^{\mathcal{G}}) \sim Q$. Here Q will be a $(m+1) \times (m+1)$ SPD matrix. Then again the regularization of co-variance matrix $\Sigma^{\mathcal{G}}$ is done as $\Sigma^{\mathcal{G}} \leftarrow \Sigma^{\mathcal{G}} + \in^{\mathcal{G}} I_m$. And then applying matrix logarithm and vectorization Q is mapped to tangent space of S^+_{m+1} by to form a $\frac{m^2+3m}{2} + 1$ dimensional feature vector denoted as $\boldsymbol{z}$.

Now we have feature vectors of all the predefined overlapping regions depicted as $\{\boldsymbol{z_g}\}_{g=1}^{G}$. We just concatenate them to form a single vector to maintain the spatial location helpful in matching the corresponding regions. Then the final representative features of a person image are depicted as $\boldsymbol{z} \ = \ [\boldsymbol{z_1^T}, \dots, \boldsymbol{z_G^T}]^T$

4.5 NORMALIZATION OF FEATURE VECTOR

The final features that we have extracted have very high dimensions due to deep structure of the model. As mentioned in (65) normalization must be performed on the high dimensional features to get better results out of them. So we have used a widely approached normalization method i.e. L2 norm normalization on our features vector as described in (61). Since we are using varying properties of an image pixel as our pixel feature vector which can have different distribution characteristics on the same image, so it may happen that some dimensions will have higher values while some very low thus some dimensions values may dominate the others while computing distance functions on it after normalization. To deal with this biasing of dimensions, we perform mean removal, prior to normalization, on training set images as in (61), defined as:

$$\boldsymbol{z} = (\boldsymbol{z} - \bar{\boldsymbol{z}}) / \| \boldsymbol{z} - \bar{\boldsymbol{z}} \|_2, \qquad 4.5\ (i)$$

where $\bar{\boldsymbol{z}}$ is the mean vector of the training set feature vectors.

CHAPTER 5

METRIC EVALUATION

In machine learning and Computer Vision, Classification is a challenge of finding to which set of categories (sub-classes) a new observation/query belongs, on the basis of a training dataset containing observations (or instances) whose category or sub-class membership is known.

The distance metric learning approach has been proposed for both unsupervised and supervised problems. For a large data set of images $\{x_i\}_{i=1}^{N} \subset \mathbb{R}^n$ in an unsupervised setting it would be expensive and tedious for a human to examine and label the entire data set. Then practically it would be better to select only a small subset of data points and examine them for getting the information on the relation between these points. Then this knowledge can be used further to classify the test data point to its correct class.

Any algorithm that performs this classification is known as a **Classifier**. The terminology "classifier" may also refer to a mathematical function, implemented by an algorithm for classification that maps input data to a known identity of class. Some of them are namely, linear classifier, quadratic classifier, Bayesian classifiers, etc.

5.1 QUADRATIC CLASSIFIER

A quadratic classifier usually implemented in the field of machine learning and classification to divide representations of many classes of elements on a quadratic surface. It is closely related to linear classifiers. Statistical form of classification considers an object as a group of vectors for all observations A and each of this objects has a known class C. This set is called a training set. Then for a given new test vector problem is to determine the closest class. Then the quadratic solution for this classifier is given as

$$A^T X A + Y^T A + Z \qquad 5.1\ (i)$$

5.1.1 Quadratic Discriminant Analysis

Linear Discriminant Analysis and Quadratic Discriminant Analysis are two efficient classifiers, with one having linear decision planes and the other constructing quadratic decision planes, respectively as shown in Fig. 8. They are efficient and used widely because they have closed-form solutions that are easy to compute, are suitable for multiclass classification, have been proved as effective in practice and have no hyper parameters to tune.

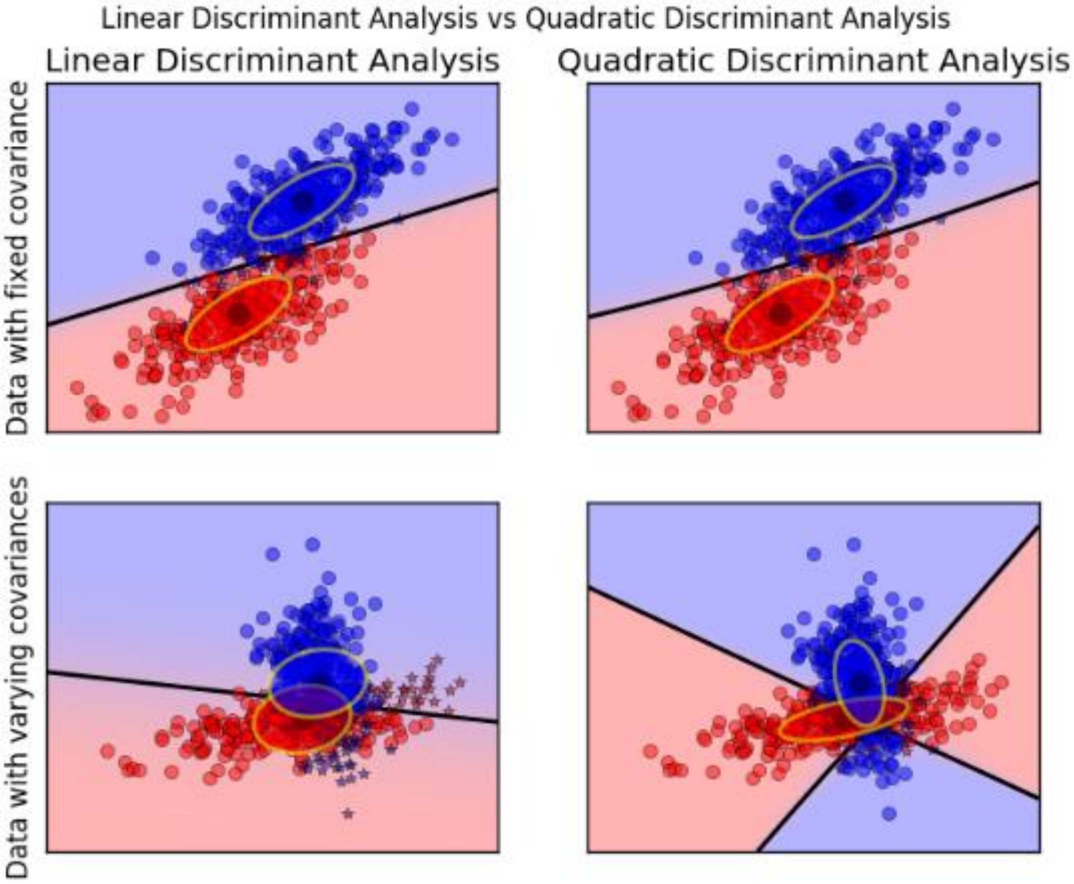


Fig. 8 LDA vs QDA

QDA can also be used to perform supervised dimensionality reduction, by projecting the training or test data to a quadratic subspace consisting of the directions which try to maximize the separation between classes. The projected subspace has dimensions necessarily less than the number of classes that proves it as a strong quality reduction. They are types of Bayesian classifiers and thus they can be derived from simple probabilistic models in which the class conditional distribution of the data points is modeled as $P(X|y = c)$ for each class c. Baye’s rule is then used to estimate the predictions:

$$P(y = c|X) = \frac{P(X|y = c)P(y=c)}{P(X)} = \frac{P(X|y = c)P(y=c)}{\sum_l P(X|y = l).P(y=l)} \qquad 5.1.1\ (i)$$

and we select the class c such that class conditional probability is maximized.

Now let’s consider that a random data point is given which was selected from class c, then the likelihood that it looked like a data point X is equivalent to the part – $P(X|y = c)$. To calculate this ‘likelihood value’, LDA and QDA use a Multivariate Gaussian Distribution model for each class.

More specifically, for linear and quadratic discriminant analysis, $P(X|y)$ is modelled as a Multivariate Gaussian distribution with density:

$$P(X \mid y = c) = \frac{1}{(2\pi)^{n}|\Sigma_c|^{1/2}} \exp(-\frac{1}{2}(X - \mu_c)^{t}\Sigma_c^{-1}(X - \mu_c)) \qquad 5.1.1(ii)$$

A probability distribution model is a way for an algorithm to understand how data points are distributed in a d-dimensional space. To utilize this function as a classifier, we just need to estimate the class priors from the training data, i.e. P(y = c) (using the proportion of events of class c), the class means μ_c (d-dimensional mean vector) and the covariance matrices Σ_c (d × d dimensional covariance matrix).They are learnt during training phase.

In the case of LDA, it is assumed that the Gaussians for each class have the same covariance matrix: $\Sigma_c = \Sigma$ for all classes c. LDA estimates separate μ_c for each class (using training points of that particular class), but Σ_c is computed for the entire training dataset. This gives linear decision planes. While for QDA, there are no assumptions on the covariance matrices Σ_c of Gaussians of different classes which leads to quadratic decision planes.

5.2 XQDA METRIC LEARNING

This was first described in LOMO (29). It is a subspace and metric learning method that learns a subspace projection matrix and a metric kernel on the subspace simultaneously. Here firstly we project the original high dimensional feature vector on the subspace and thus the dimensionality of the feature vector is reduced. Then using Mahalanobis distance metric on the reduced dimension feature vector we calculate the similarity score between the probe and gallery images.

XQDA (29) is largely an extension of Bayesian face and KISSME, as in this algorithm a discriminant subspace is further learned along with a metric kernel. So first let us revise these methods in short.

5.2.1 RELATED WORKS REVISIT

Work summary as in (29), denoted the samples by vector X and their corresponding classes by vector Y. Consider a sample difference $\lambda = x_i - x_j$. If $y_i = y_j$ then λ will be called as intrapersonal difference while it will be called the extra-personal. difference if $y_i \neq y_j$ (59). Accordingly, two categories of variations can be defined: the intra class variations Ψ_I and the external class variations Ψ_E. Therefore, by distinguishing the above two classes our multi-class classification problem can be easily solved. Moghaddam et al. (59) proposed that each of the two classes be modelled with a Multivariate Gaussian distribution. This will lead us to a QDA model with Ψ_I and Ψ_E defined as our two classes. Furthermore, looking into (59) tells us that both Ψ_I and Ψ_E have zero mean. This algorithm was called Bayesian face and was applied to face recognition. Interestingly, in (25), a similar approach called KISSME was derived via the log likelihood ratio test of the two Gaussian distributions, and was applied to the problem of person re-identification.

In (29) its given that for both of these techniques under the zero-mean Gaussian distribution the likelihoods of observing λ in Ψ_I and Ψ_E are defined as:

$$P(\lambda|\Psi_\mathrm{I}) = \frac{1}{(2\pi)^{d/2}|\Sigma_I|^{1/2}} \exp(-\frac{1}{2}\lambda^T \Sigma_I^{-1}\lambda) \qquad 5.2.1\text{ (i)}$$

$$P(\lambda|\Psi_\mathrm{E}) = \frac{1}{(2\pi)^{d/2}|\Sigma_E|^{1/2}} \exp(-\frac{1}{2}\lambda^T \Sigma_E^{-1}\lambda) \qquad 5.2.1\text{ (ii)}$$

where $\Sigma_I \; and \; \Sigma_E$ are the co-variance matrices of Ψ_I and Ψ_E, respectively, and $\eta_I \; and \; \eta_E$ tells the number of samples in the two classes. By applying the Bayesian rule and the log-likelihood ratio test (29), the decision function can be simplified as:

$$f(\lambda) = \lambda^T(\Sigma_I^{-1} - \Sigma_E^{-1})\lambda \qquad 5.2.1\text{ (iii)}$$

and so the distance function is derived between $x_i \; and \; x_j$ as:

$$d(x_i, x_j) = \left(x_i - x_j\right)^T (\Sigma_I^{-1} - \Sigma_E^{-1})(x_i - x_j) \qquad 5.2.1\text{ (iv)}$$

This is equivalent to estimating covariance matrices $\Sigma_I \; and \; \Sigma_E$

5.2.2 XQDA METRIC

As we know that our final feature dimensions d is large, and for classification this will require high computational capabilities of the system. So, a low dimensional space $\mathbb{R}^r(r < d)$ is more suitable. In Bayesian Face method (59) it was suggested that $\Sigma_I \; and \; \Sigma_E$ be decomposed separately so as to reduce the feature dimensions. In KISSME algorithm (25), dimensionality reduction was first done by PCA and then $\Sigma_I \; and \; \Sigma_E$ were estimated in the subspace of PCA. However, in both methods the dimension reduction does not care for the distance metric learning and thus they are not optimal.

So in LOMO (29), Bayesian face and KISSME algorthims were extended to cross-view metric learning, where a subspace was learnt, represented as $W = (w_1, w_2, \dots, w_r) \in \mathbb{R}^{d \times r}$, with cross-view data. It also learnt a distance metric kernel in r-dimensional subspace, for getting similarity measure between the cross-view data, simultaneously. *Liao. et. al* (29) represented a cross-view training set as $\{X, Z\}$ with total c classes. $X = (x_1, x_2, \dots, x_n) \in \mathbb{R}^{d \times n}$ containing n samples in a d-dimensional space from one view and $Z = (z_1, z_2, \dots, z_m) \in \mathbb{R}^{d \times m}$ containing m samples in the same d-dimensional space but from the other view. The distance function given in Bayesian face method when computed in the r dimensional subspace W (29) is defined as:

$$d_W(x, z) = (x - z)^T W(\Sigma_I'^{-1} - \Sigma_E'^{-1}) W^T (x - z) \qquad 5.2.2\ (i)$$

where $\Sigma_I' = W^T \Sigma_I W$ and $\Sigma_E' = W^T \Sigma_E W$. Then the kernel matrix is learnt as:

$$M(W) = W(\Sigma_I'^{-1} - \Sigma_E'^{-1})\, W^T \qquad 5.2.2\ (ii)$$

Now we need to optimize d_W, but doing this directly is difficult because the subspace matrix W is constrained by the two inverse matrices.

It was told in previous works (59) that Ψ_{I} and Ψ_{E} have zero mean. So the projected samples of the two classes, on a given basis w, (29) will also centre at zero. But they may have variances σ_I *and* σ_E differing from each other and thus can be used for distinguishing two classes as shown in Fig. 9. Therefore, in the method being used in our work we try to maximize $\sigma_E(w)/\sigma_I(w)$ (objective function) by optimizing the projection direction w. Here $\sigma_I(w) = w^T \Sigma_I w$ and $\sigma_E(w) = w^T \Sigma_E w$, therefore as described in (29) the objective function $\sigma_E(w)/\sigma_I(w)$ is equivalent to Generalized Rayleigh Quotient

$$J(w) = \frac{w^T \Sigma_E w}{w^T \Sigma_I w} \qquad 5.2.2\ (iii)$$

Maximization of $J(w)$ can be defined as

$$\max_w w^T\, \Sigma_E w, s.t.\ w^T \Sigma_I w = 1 \qquad 5.2.2\ (iv)$$

Then using the Generalized Eigen-value Decomposition problem as used in LDA this can be solved. According to this problem (29), of all the eigenvalues of $\Sigma_I^{-1} \Sigma_E$ the largest one is the maximum value of $J(w)$, and the corresponding eigenvector w_1 is our solution. The eigenvector corresponding to the second largest eigenvalue gives the second largest value of $J(w)$ and is also orthogonal to w_1. Then our discriminant subspace is given as $W = (w_1, w_2, \dots, w_r)$ and a kernel matrix is also learnt. This kernel is then used in the distance function applied on the learned subspace. We have used Mahalanobis distance function to get the similarity scores between the probe and gallery set images.

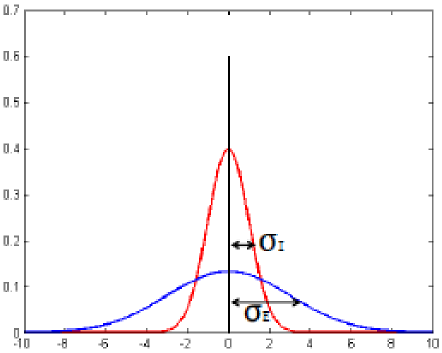


Fig. 9 Distributions of Intra Class (I) and External Class (E) Gaussian in one projected dimension

5.2.3 MAHALANOBIS DISTANCE METRIC

It often happens in many machine learning problems that the Euclidean distance functions applied between data points may not present the topology that we are trying to capture. Then Kernel methods come to the rescue and address this problem by mapping the input data points into new spaces where the Euclidean distance function can be applied. But there is an alternative approach that has been used widely in recent researches, i.e. to construct a Mahalanobis distance (quadratic Gaussian metric) over the original input space instead of using Euclidean distances. Recently a lot of interest has been taken in learning a kernel matrix function for getting the distance or similarity metric based on the class of Mahalanobis distance functions. In general, a Mahalanobis distance metric tries to measure the distance between the square of two data points $x_i \: and \: x_j$:

$$D_M\left(x_{ik}, x_{jk}\right) = \left(x_i - x_j\right)^T M\left(x_i - x_j\right) \qquad 5.2.3 \text{ (i)}$$

where $M \geq 0$ is a kernel learnt during the training phase and is a positive semi-definite matrix and $x_i, x_j \in \mathbb{R}^d$ is a pair of samples (i, j).

So using this function we calculate the similarity or distance score between the feature vectors of the probe set of images and the gallery set in the projected subspace. Then we sort these score values in ascending order so as to get the closest or most similar match at the top of the ranking scale. Using these sorted scores we can then retrieve all the occurrences of a person in other camera views.

CHAPTER 6

EXPERIMENT

To show the efficiency of our method we have performed various analysis and comparisons with other feature descriptors and metrics and have showed the results in the form of CMC curves and Ranking tables. We have also performed image retrieval experiment on some of the datasets considering person re-identification as a recognition problem.

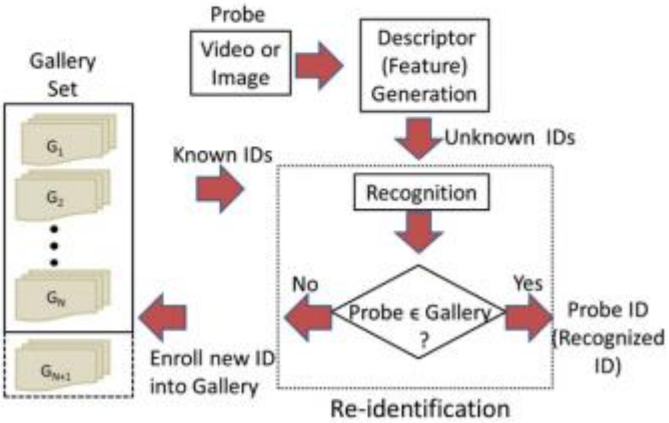


Figure 10: Re-Id as a Recognition problem

6.1 DATASETS AND DATABASE INFORMATION

As we know that various challenges occur while acquiring the dataset of images across different cameras which change the visual characteristics of a person. Some of them are illumination variations, poses, view angles, scales and camera resolutions. Further there may be occlusion by objects or people and cluttered. Thus, in order to analyse and compare the robustness of Re-ID techniques it is important to acquire data that inherits these factors. We have used four datasets for the evaluation of the method’s effectiveness, namely VIPeR, CUHK01, PRID450S and GRID.

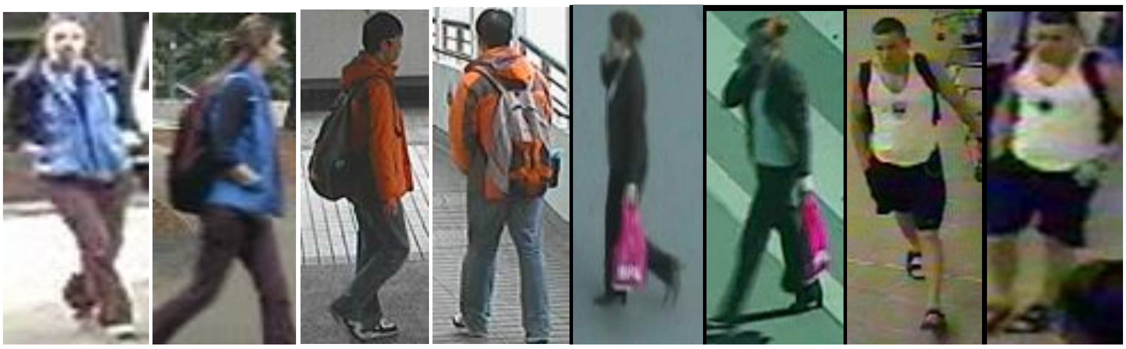
For evaluation database of each dataset is created that has the following fields:

- allimagenames -- all image names of the dataset
- traininds_set/testinds_set -- index of the training/test images for each division (index order is the same as allimagenames)
- trainimagenames_set/testimagenames_set -- image names of the each training/test division
- trainlabels_set/testlabels_set -- person IDs of the training/test images for each division
- traincamIDs_set/testcamIDs_set -- camera IDs of the training/test images for each division

6.1.1 VIPER

The VIPeR dataset (5) has total 1264 images of 632 individuals captured from two different camera views Camera A and Camera B. Each individual has a single image in each camera. This dataset was collected to test viewpoint invariant pedestrian recognition and hence different viewpoints were captured. The view angles were roughly quantized into 45° angles. The dataset also has illumination variations between image pairs. The images are cropped and resized to 128 × 48 pixels.

For evaluation using training and testing phases the set of 632 image pairs is randomly split into two sets, one for training and other for testing with 316 image pairs in each and 10 such sets are created. For each image pair, Image from one camera is assigned as probe and the other camera as gallery, randomly. This is done in both training and testing phase. Then the process of selecting a single image from the probe set and matching it with all images from the gallery set is then repeated for all images in the probe set.



(a)VIPeR (b) CUHK01 (c) PRID450s (d) GRID

Fig. 11 Dataset Image pairs: Pairs of images of the same person taken from different cameras, from four benchmark datasets

6.1.2 CUHK01

The Campus dataset (66) has 3884 images in total of 971 individuals manually cropped to 60x160 pixels. Two disjoint camera view were selected where camera A captures more pose and viewpoint variations while camera B captures mainly frontal view and back view. Each person has two images captured in each camera.

Training set contains 486 while testing set contains 485 individuals selected randomly and further they are divided into probe and gallery set same as we did in the case of VIPeR dataset.

6.1.3 PRID450S

The PRID 450S dataset (23) is taken from PRID 2011, but then it re ordered as pair of images in two different camera A and B as done in VIPeR. It contains 450 single-frame image pairs of walking humans taken in two spatially disjoint camera views. Then the patches which contained the person were segmented manually by boundaries of resolution 100-150 pixels from original content in which the resolution was 720×576 pixels.

In this we have assigned 225 individuals each to training and testing set selected randomly and separated as probe and gallery sets in the same manner as above.

6.1.4 GRID

The QMUL underground Re-Id (GRID) (67) dataset contains 250 pedestrian image pairs. For every single individual we have two frames, each captured in different camera view in a busy underground station. It has captured the challenges of variations of pose, colors, lighting changes; as well as poor image quality caused by low spatial resolution very well. Two folders are provided namely ‘Probe’ and ‘Gallery’ each containing 250 images of 250 individuals while gallery folder also contains additional 775 images that do not belong to probe set. These images are kept fixed in the testing set during cross validation.

Each image in all these datasets is resized to 128×48 pixels to provide equal grounds for analyzing the descriptor.

6.2 FEATURES SETTING

We have used four different sets of d-dimensional pixel level feature vector

Pixel Feature 1 (YCM) - In this we have used y distance and first two color moments i.e. mean and standard deviation of RGB color space to create a 7-d pixel feature. Color moments are calculated on patches of size 5×5 with patch interval of size 2 to reduce the computational complexity. Thus it reduces the size of the image to 64×24. It can be represented as - $[y, mean(R), std(R), mean(G), std(G), mean(B), std(B)]$.

Pixel Feature 2 (SCHMID) – This is a texture feature representation of pixels comprising of 13-d Schmid filter responses using Schmid filter banks applied on 10×10 non-overlapping patches. Banks parameters (σ, τ) pairs for 13 filters are (2,1), (4,1), (4,2), (6,1), (6,2), (6,3), (8,1), (8,2), (8,3), (10,1), (10,2), (10,3) and (10,4).

Pixel Feature 3 (YGOHSV) – Then we have used a combination of 1-d y distance value, 4-d gradient orientation and 3-d HSV color space values; in total 8-d pixel feature representation.

Pixel Feature 4 (YGOnRnG) – It is a 7-d pixel feature representation comprising of a combination of 1-d y distance value, 4-d gradient orientation and 2-d nRnG color space values. Here the nRGB is the normalized color space obtained by normalizing RGB color space values (*e.g.*, nR = R/(R+G+B)). We have used only {nR, nG} values, since this color space has redundancy.

After constructing d-dimensional pixel feature we have extracted multi-level Gaussian descriptor for seven horizontal strips or regions (R=7) that are overlapping. Each strip comprises 32×48 pixels. In each region local patch Gaussians are extracted from the patches of 5×5 pixels with patch interval 2.

6.3 PERFORMANCE ANALYSIS AND COMPARISON

We have the evaluated our procedure on 10 sets of random data splits into training and testing sets. We have represented the results in the form of average Cumulative Matching Characteristic (CMC) curves and comparisons of recognition percentage of ranks in tabular form. The training data is divided uniformly into a gallery and probe set. Then every pair of image is further divided and then assigned, in no specific rule, to probe and gallery set. Then CMC curve is generated for the query set by selecting a query image (image from camera a/camera b) and sorting its similarity scores to the gallery images (i.e. every image in camera b/camera a). It gives the rank for every image in the gallery, relative to the selected probe image. This procedure is done in repetition for each image in probe set and averaged. Thus the CMC curve is then the expectation of getting the true match in first r matches.

6.3.1 Comparison Of Different Pixel Features

We have shown the effectiveness of each pixel feature separately and also by combining them on CMC curves as shown in Fig. 12 (a), (b), (c). In the Table 3 we have presented rank-1, rank-10 and rank-20 recognition rates. Clearly rank-1 recognition rate of Y-Color Moment pixel feature is highly greater than that of Schmid filter pixel feature. It proves that for person re-id problem color features perform much better than texture features. But texture feature when combined with color feature exceeds the rank-1 recognition rates of the color feature when used alone. We have also showed the identification rates for different combination of pixel features in Table 3.

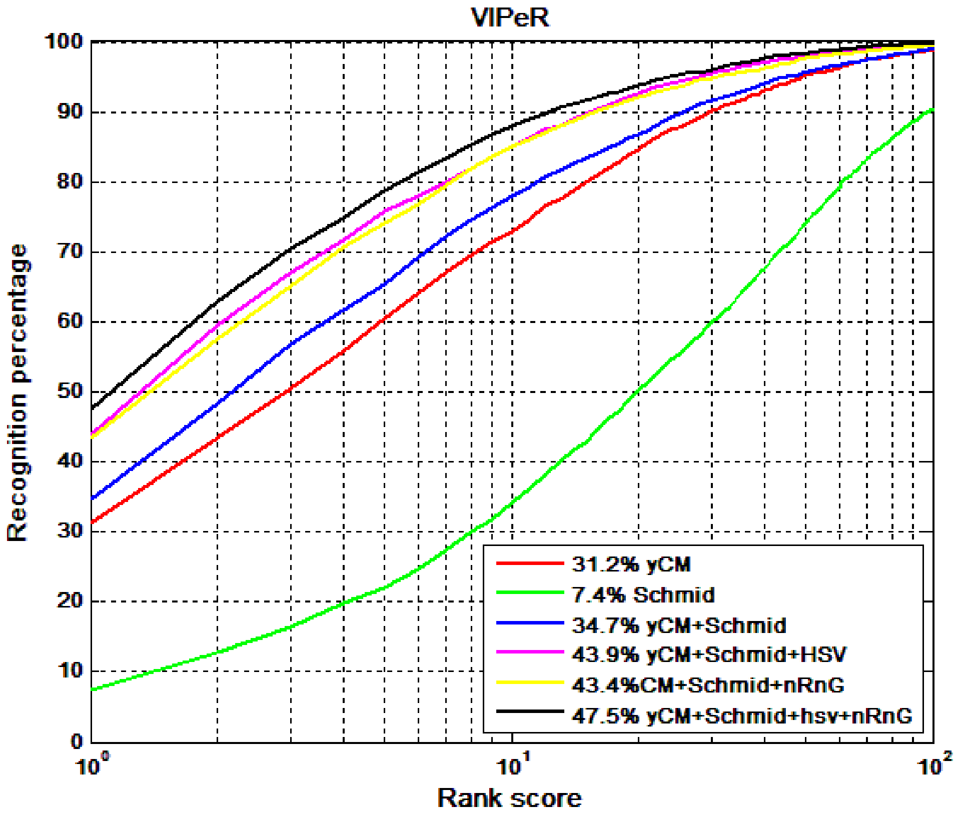


Fig. 12(a) CMC curves for VIPeR dataset

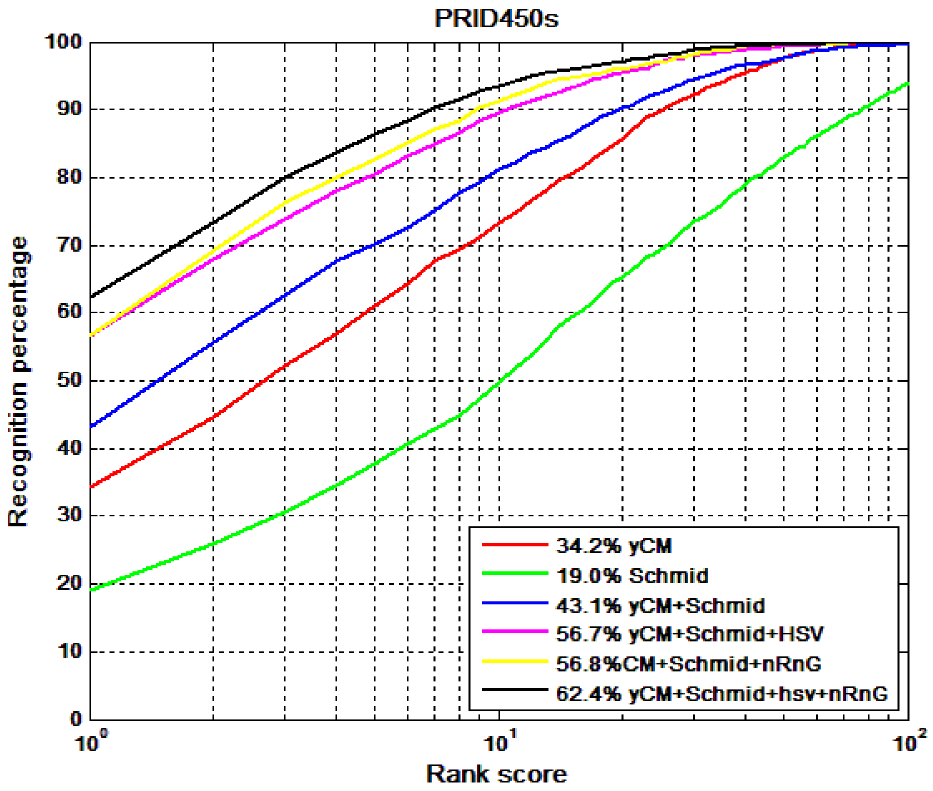


Fig. 12(b) CMC curves for PRID450s dataset

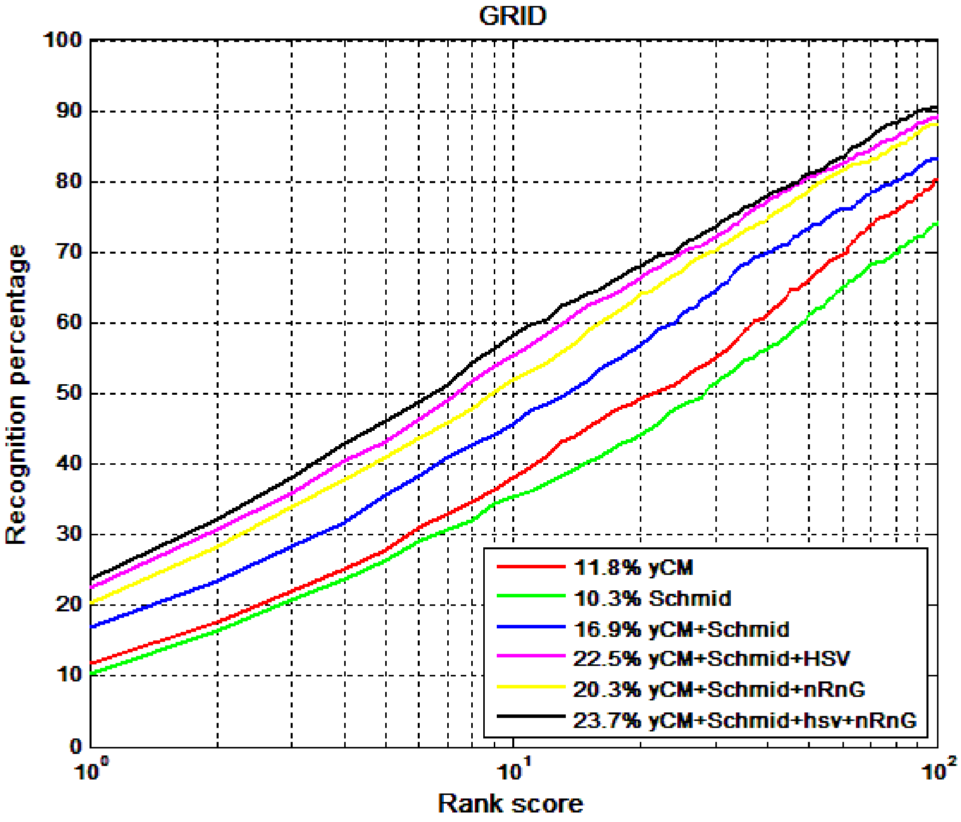


Fig. 12(c) CMC curves for GRID dataset

Table 3: Comparison of recognition rates for different combination of pixel features on VIPeR, PRID450s and GRID dataset. The best results are highlighted by bold numbers.

Pixel Feature Used	VIPeR			PRID450s			GRID		
	R=1	R=10	R=20	R=1	R=10	R=20	R=1	R=10	R=20
Y+Color moment(YCM)	31.2	72.9	86.8	34.2	73.3	85.7	11.8	38.2	49.2
Schmid	7.4	34.2	50.1	19.0	49.8	65.3	10.3	35.4	44.1
YCM+Schmid	34.7	78.1	86.8	43.1	81.1	90.3	16.9	45.7	56.7
YCM+Schmid+HSV	43.9	85.1	92.5	56.7	89.6	95.4	22.5	55.3	66.4
YCM+Schmid+nRnG	43.4	85.1	92.0	56.8	91.2	96.0	20.3	52.0	64.2
YCM+Schmid+HSV+nRnG	**47.5**	**87.9**	**93.7**	**62.4**	**93.5**	**96.9**	**23.7**	**58.2**	**68.1**

6.3.2 Comparison with Other Meta Descriptors

We have compared our multi-level Gaussian distribution model with covariance distribution models. Covariance descriptor (19) when used globally on the image gives 26.9 % rank-1 rates on VIPeR dataset. The Cov-of-Cov is also a hierarchical descriptor which uses only covariance modelling of both patch and region (39) (38). It gives 33.9% rank-1 rates on VIPeR approximately 7 % better than non-hierarchal single level covariance rates.

We also compare our descriptor with other single-layered meta descriptors as chosen in (61) namely: Heterogeneous Auto-Similarities of Characteristics (HASC) (68), Hybrid Spatiogram and Covariance Descriptor (HSCD) (20), Local Descriptors encoded by Fisher Vector (LDFV) (16), Second-order Average Pooling (2AvgP) (33) and GOLD (37) .

The HASC (68) is a fusion descriptor of the covariance and the Entropy and Mutual Information (EMI) descriptor. It encodes linear relations using co variances while nonlinear associations are encoded by information-theoretic measures like mutual information and entropy. Both of these descriptors have equal dimensions and EMI descriptor tries to capture the non-linear dependency within pixel features. The HSCD is again a hybrid descriptor composed of spatiogram and covariance feature. Spatial histograms of different regions are accumulated and three sub features are then extracted in spatiogram. While in covariance feature, several color spaces and intensity gradients are taken as pixel features and then statistical feature vectors are extracted from pyramid of covariance matrices.

In LDFV pixel features are encoded using Fisher Vector coding. This coding tries to encode difference of pixel features from GMM means which are pre-trained. We have set the number of GMM components to 16 as recommended in (16). The GOLD (37)uses mean vector and covariance matrix to describe a region of image and also applies Log-Euclidean metric and half vectorization to flatten the covariance matrix. The 2AvgP (33) also describes an image region by the zero-mean covariance matrix using second-order generalizations of average and max-pooling, and applies LEM and half-vectorization to obtain a robust feature vector.

We have listed the performance of our method in Table 4(a) and 4(b)and the compared descriptors. All these descriptors in Table 4(b) are non-hierarchical or single layered descriptors and ignore the local properties of regions. They have similar performances. We can clearly see that our descriptor has good rank-1 rate of 47.5 % that is outperforming all these methods. It is also evident from the table that the both the hierarchical methods i.e., our method and the one given by (38) cov-of-cov, have higher rank-1 matching rates than the non-hierarchical ones and thus it can be inferred that hierarchical model always performs better. Between the two hierarchical methods mentioned, the difference is that we have used mean information together with the covariance value while (38) has used just the covariance information. Since our method has 13.6 % better rank-1 rates than covariance hierarchical descriptors so this proves that mean information is valuable feature of the image.

Table 4: Comparison of our descriptor with other meta descriptors wirh best results highlighted in bold

	Descriptors	VIPeR			PRID450s			GRID			CUHK01		
		R=1	**R=10**	**R=20**	**R=1**	**R=10**	**R=20**	**R=1**	**R=10**	**R=20**	**R=1**	**R=10**	**R=20**
(a)	MLGD (Our method)	**47.5**	**87.9**	**93.7**	**62.4**	**93.5**	**96.9**	**23.7**	**58.2**	**68.1**	**54.5**	**83.5**	**90.5**
	COV-of-COV (38)	33.9	76.6	87.7	47.0	83.4	91.6	16.6	45.0	55.2	40.9	72.5	81.1
(b)	HASC (68)	30.9	70.6	81.8	41.8	76.3	85.2	12.9	35.6	47.3	38.6	68.7	77.1
	GOLD (37)	27.1	66.5	77.7	40.5	73.8	82.2	10.9	29.2	37.4	35.3	65.2	74.2
	HSCD (20)	31.2	86.5	91.8	-	-	-	-	-	-	-	-	-
	2AvgP (33)	28.8	68.5	79.2	44.7	75.8	83.8	12.9	36.7	47.4	36.1	68.1	76.3
	LDFV (16)	25.3	66.8	79.4	32.1	66.9	77.6	16.2	41.9	53.1	36.4	71.0	80.3
	Cov (19)	26.9	65.8	77.1	40.4	73.4	82.1	10.6	29.0	36.7	34.5	64.5	73.6

6.3.3 Comparison with Descriptors Using XQDA Metric

We have also shown that despite of using the same metric learning method our descriptor is also better than some other descriptors using the same XQDA metric. This idea to apply the XQDA metric on them comes from (61), and the descriptors used are also same, namely, LOMO (29), Color Histogram + LBP (69) and gBiCov (15). To compare them on different datasets their features these features were extracted on these datasets using the available source codes

We can clearly see from Table 5 that, excluding PRID450s dataset, in rest all, the rank-1 matching rates of our Gaussian fusion method exceed all other descriptor in matching. LOMO feature rates are the second highest in the comparison table. So we can infer from this that since the classification step in all the cases has used a common metric so the difference in rates is all due to descriptor performance. Thus our descriptor clearly has some robust qualities.

Table 5: Comparison of our descriptor with other descriptors using the same XQDA metric for all.

Descriptors	VIPeR			PRID450s			GRID			CUHK01		
	R=1	**R=10**	**R=20**	**R=1**	**R=10**	**R=20**	**R=1**	**R=10**	**R=20**	**R=1**	**R=10**	**R=20**
MLGD(Our method)	**47.5**	**87.9**	**93.7**	62.4	**93.5**	**96.9**	**23.7**	**58.2**	**68.1**	**54.5**	83.5	90.5
LOMO (29)	41.1	82.2	91.1	62.6	92.0	96.6	17.9	46.3	56.2	49.2	84.2	90.8

CH+LBP (69)	27.7 69.3 82.4	21.5 60.8 74.4	16.2 45.0 57.1	31.3 70.4 81.5
gBiCOV (15)	22.8 64.0 77.8	27.9 67.2 76.8	10.6 30.4 41.4	24.1 55.6 67.2

6.3.4 Comparison With Some Popular State-Of-The-Art Methods

In Table 6, we have listed the reported results on some of the state-of-the-art methods, including Metric Ensemble (69), Mid Level Filter Learning (MLFL) (18), SCNCD (70), Salience Matching (17), Semantic attribute representation (71) and LOMO (29). It is clearly visible that our descriptor performance goes above many state-of-the-art methods and creates new state-of-the-art results, i.e., 47.5%, 62.4%, 54.5% and 23.7% rank-1 rates on VIPeR, PRID450S, CUHK01 and GRID dataset, respectively. Since LOMO and our method make use of the common metric learning method, it is evident that our feature descriptor has better design and recognition capabilities. The proposed descriptor also outperforms the efficient metric ensemble (69) by 1.6 % rank-1 rates.

Table 6: Comparison with well known State-of-art results. The best results are highlighted with bold font.

Descriptors	**VIPeR**				**CUHK01**			
	R=1	**R=5**	**R=10**	**R=20**	**R=1**	**R=5**	**R=10**	**R=20**
MLGD(Our method) + XQDA	**47.5**	**78.8**	87.9	93.7	**54.5**	75.7	83.5	90.5
Metric Ensemble (70)	45.9	77.5	**88.9**	**95.8**	53.4	**76.4**	**84.4**	90.5
LOMO + XQDA (29)	40.0	-	80.5	91.1	49.2	75.7	84.2	**90.8**
SCNCD (71)	37.8	68.5	81.2	90.4	-	-	-	-
Semantic (72)	31.1	68.6	82.8	94.9	32.7	51.2	64.4	76.3
SalMatch (17)	30.2	52.0	65	-	28.5	45.0	55.0	-
MLFL (18)	29.1	-	65.9	70.9	34.3	55	65	75

6.3.5 Running time

We have evaluated a working model for the system by coding on Matlab. We have used MEX functions given by (61) for calculating the covariance matrices. The system used for computations is a PC with Intel(R) Core(TM) i5-2450M @ 2.50GHz CPU. The total evaluation time for single feature descriptor and as well as the fusion case is as shown in Table 6. The time displayed represents the average time taken for all the images in a dataset. Results in Table 6 correspond to the images of VIPeR dataset. The proposed descriptor is computationally expensive and slower than covariance descriptor while it is 5 times faster than BiCov[29]. The recognition cost of our method is nearly equal to LOMO while descriptor is slower than LOMO. But there are methods like (18) (17) (14) which have even higher computational cost compared to our method and thus our method still ought to be approached for.

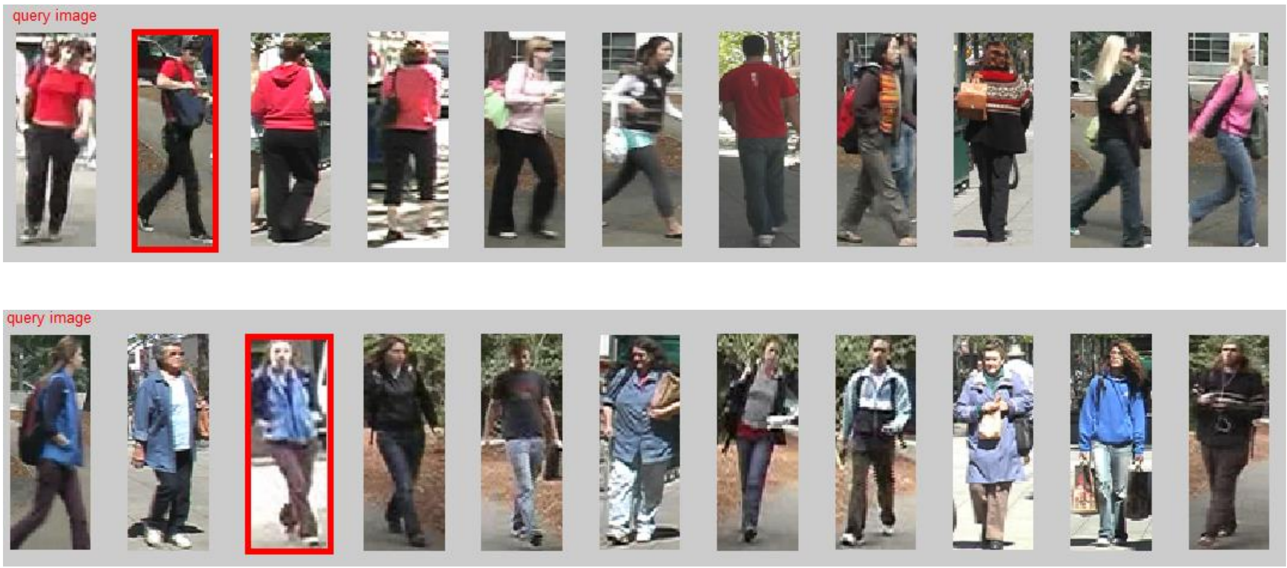
Table 7: Average feature extraction time (seconds/image).

Cov(for RGB pixel features)	YCM	Schmid	Fusion MLGD	LOMO	gBiCov
0.021	1.451	2.821	5.21	0.016	7.8

6.4 IMAGE RETRIEVAL RESULTS

We have also performed an image retrieval experiment where we have tried to retrieve the true match of a probe image from the disjoint camera images. We have performed this on three datasets namely, VIPeR, CUHK01and PRID450s and displayed the results in the Fig 13.

We have first given a probe image to the system and then retrieved closest 10 images from the disjoint gallery set. One of these 10 retrieved images contain the true match of the probe image from a different view angle and is highlighted by red boxes in the Fig 13. It must be mentioned here that while VIPeR and PRID450s have singleframe per person in each camera images, CUHK01 has two or more image frames per person in each camera images.



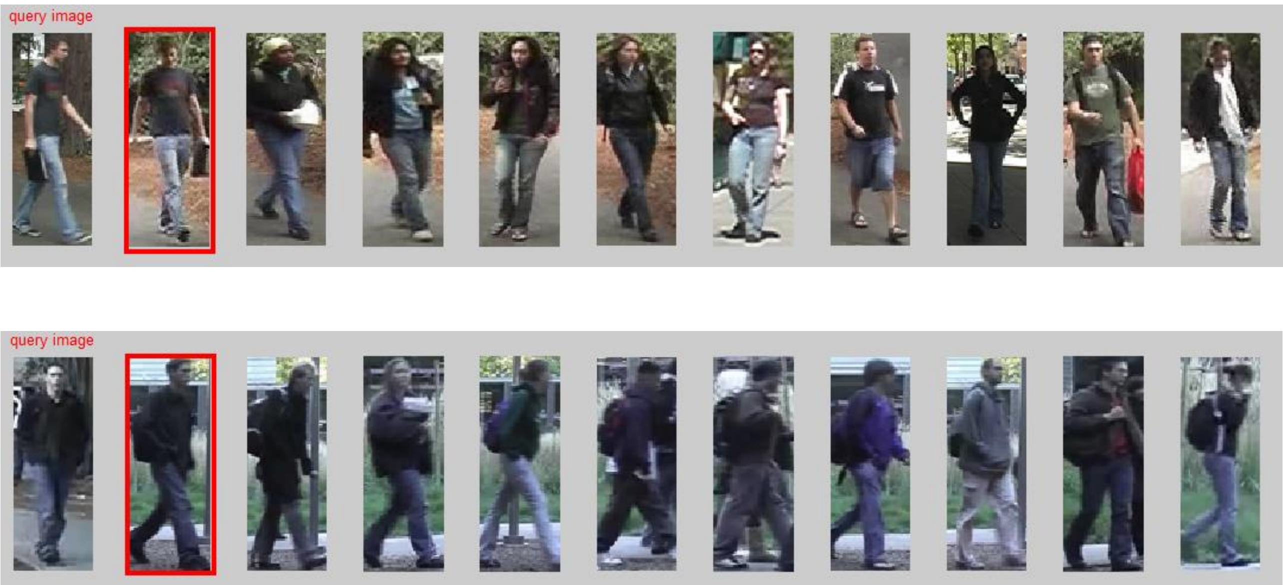


Fig. 13(a) Image retrieval results on VIPeR

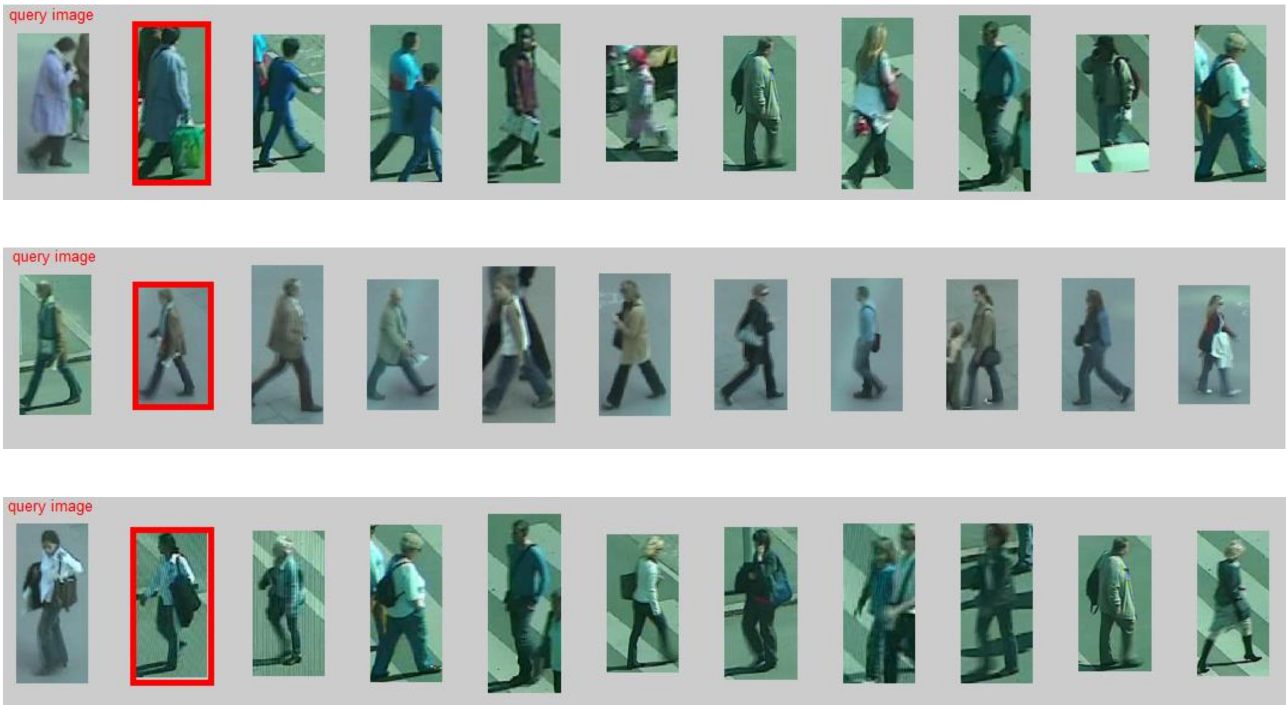


Fig. 13(b) Image retrieval results on PRID450s

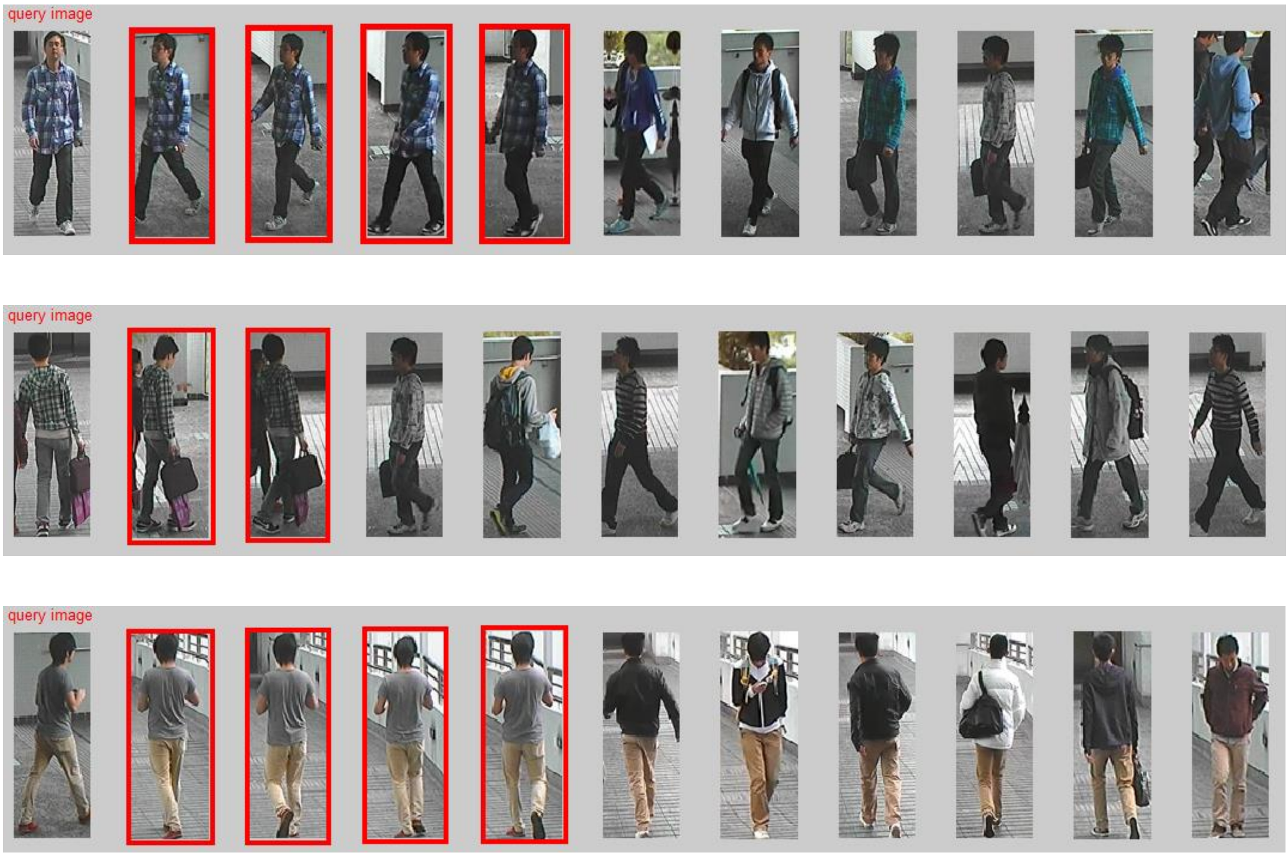


Fig 13(c) Image retrieval results on CUHK01

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

This work presents an efficient and effective method, a multi-layer Gaussian descriptor model, for the problem of person re-ID. The descriptor presented in this work utilizes both value of mean and covariance of pixel features present in an image and thus returns a robust and discriminative representation of data. It first models the local patches as Gaussian distribution and then from a set of local patches it models the region descriptor as Gaussian. In this way it does not neglect the relevant information in the local structures of the image while acquiring a global representation. The results of our in depth experiments proved that the proposed descriptor can even outperform the state-of-the-art performances on four public datasets. We have also conducted image retrieval experiment where we treat this problem as a recognition problem. The results prove that our descriptor is robust against viewpoint changes and illumination variations. For effective metric learning we have used XQDA metric described in LOMO (29). It is formulated as a Generalized Rayleigh Quotient, and by applying generalized eigenvalue decomposition a closed-form solution can be obtained.

In future the deep network of Gaussian descriptors can be explored to consider the local structure of human appearances in more depth. In addition, some other kinds of pixel features can be tested to search for any further improvements in the process accuracies. We can try to find computationally low cost features to reduce the extraction and search time of the system. Then we can also try to learn a new metric which is highly discriminative for high dimensional feature vectors.

BIBLIOGRAPHY

[1] **M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani,** "*Person re-identification by symmetry-driven accumulation of local features"*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco , CA, pp. 2360-2367, 2010

[2] **D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino,** "*Custom Pictorial Structures for Re-identification"*, 2011. Proc. British Machine Vision Conference, BMVA Press, pp. 68.1-68.11, 2011

[3] **S. Bak, F. Bremond, E. Corvee, M. Thonnat,** "*Person Re-identification Using Spatial Covariance Regions of Human Body Parts",*, 2010 IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, pp. 435-440, 2010

[4] **S. Bak, F. Bremond, E. Corvee, M. Thonnat,** "*Multiple-shot human re-identification by Mean Riemannian Covariance Grid"*, 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Klagenfurt, pp. 179-184, 2011

[5] **Gray, D. and Tao, H,** *"Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features"*, 2008. Proc. 10th European Conference on Computer Vision, Springer, pp. 262-275, 2008

[6] **S. Pedagadi, J. Orwell, S. Velastin and B. Boghossian,** "*Local Fisher Discriminant Analysis for Pedestrian Re-identification"*, 2013. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, pp. 3318-3325, 2013

[7] **Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao and J. R. Smith,** "*Learning Locally-Adaptive Decision Functions for Person Verification"*, 2013. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR , pp. 3610-3617, 2013

[8] **O. Oreifej, R. Mehran, M. Shah,** "*Human Identity Recognition in Aerial Images.* 2010. Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 709–716, 2010

[9] **Y. Zhang and S. Li,** *"Gabor-LBP Based Region Covariance Descriptor for Person Re-identification"*, 2011. Sixth International Conference on Image and Graphics, pp. 368-371, 2011

[10] **Tuzel, Oncel, Porikli, Faith and Meer, Peter,** "*Pedestrian Detection via Classification on Riemannian Manifolds."*, 2008, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 30, no. 10, pp. 1713-1727.2008

[11] **Forssen, P. E.,** "*Maximally Stable Colour Regions for Recognition and Matching",* 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minnepolis MN, pp. 1-8, 2007

[12] **M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja,** *"Pedestrian Recognition with a Learned Metric",* 2010 Proc. 10th Asian conference on Computer vision (ACCV), Queenstown, New Zealand Springer pp.501-512, 2010

[13] **B. Alipanahi, M. Biggs, A. Ghodsi, et al.,** "*Distance metric learning vs. fisher discriminant analysis.* 2008. International conference on Artificial intelligence,Vol. 2, pp. 598-603, 2008

[14] **Bazzani, L., Murino, V. and Cristani, M.,** "*Symmetry-driven accumulation of local features for human characterization and re-identification",* 2013 Computer Vision and Image Understanding, , Vol. 117, pp. 130-144, 2013

[15] **B. Ma, Y. Su, and F. Jurie**, *"Covariance descriptor based on bio-inspired features for person re-identification and face verification",* 2014, Image and Vision Computing, Vol. 32, pp. 379–390, 2014

[16] **B. Ma, Y. Su, and F. Jurie,** "*Local descriptors encoded by fisher vectors for person re-identification",* 2012. European Conference on Computer Vision Workshops, Florence, Italy, Vol. I, pp. 413-422, 2012

[17] **R. Zhao, W. Ouyang, and X. Wang,** "*Person re-identification by salience matching",* 2013. IEEE International Conference on Computer Vision, Sydney, NSW, pp. 2528-2535.. 2013

[18] **R. Zhao, W. Ouyang and X. Wang,** "*Learning Mid-level Filters for Person Re-identification",* 2014 Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 144-151, 2014

[19] **O. Tuzel, F. Porikli, and P. Meer,** *"Region covariance: A fast descriptor for detection and classification".* 2006 European Conference on Computer Vision (ECCV). pp. 589-600, 2006

[20] **Mingyong Zeng, Z. Wu, C. Tian, Lei Zhang and Lei Hu.,** "*Efficient person re-identification by hybrid spatiogram and covariance descriptor",* 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA , pp. 48-56, 2015

[21] **S. Bak, G. Charpiat, E. Corv´ee, F. Br´emond, and M. Thonnat,** "*Learning to match appearances by correlations in a covariance metric space", 2012*. European Conference on ComputerVision (ECCV), pp. 806-820, 2012

[22] **S. Bak, E. Corv´ee, F. Br´emond, and M. Thonnat,** *"Boosted human re-identification using Riemannian manifold",* 2012 Image Vision and Computing, Vol. 30, pp. 443–452, 2012

[23] **P. M. Roth, M. Hirzer, M. K¨ostinger, C. Beleznai, and H. Bischof.,** "Mahalanobis distance learning for person reidentification", 2014 *Person Re-Identification,* pp. 247-267, 2014

[24] **J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon.,** "*Information-theoretic metric learning",.* 2007 Proc. IEEE International Conference on Machine Learning,, Corvalis, USA, pp. 209-216, 2007

[25] **M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth,H. Bischof.,** "*Large scale metric learning from equivalence constraints.",* 2012 In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288-2295, 2012

[26] **K.Weinberger, J. Blitzer, and L. Saul.,** "*Distance metric learning for large margin nearest neighbor classification",* 2009 The Journal of Machine Learning Research, Vol. 10, pp. 207-244, 2009

[27] **W.S. Zheng, S. Gong, and T. Xiang***, "Person re-identification by probabilistic relative distance comparison",* 2011. ComputerVision and Pattern Recognition, pp. 649-656, 2011

[28] **M. Hirzer, P. M. Roth, M. K¨ostinger, and H. Bischof.,** "*Relaxed pairwise learned metric for person re-identification",* 2012. European Conference on Computer Vision, Florence, Italy, pp. 780-793, 2012

[29] **S. Liao, Y. Hu, X. Zhu, and S. Z. Li.,** "*Person re-identification by local maximal occurrence representation and metric learning",* 2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197-2206, 2015

[30] **Nagaoka, S. Amari and H.,** " *Methods of Information Geometry",* 2001 Translations of mathematical monographs, Vol. 191.

[31] **V. Arsigny, P. Fillard, X. Pennec, and N. Ayache.,** " *Geometric means in a novel vector space structure on symmetric positive-definite matrices",* 2006, SIAM J. Matrix Analysis Applications, Vol. 29, pp. 328–347, 2006

[32] **P. Li, Q. Wang and L. Zhang,** *"A Novel Earth Mover's Distance Methodology for Image Matching with Gaussian Mixture Models"*, 2013. IEEE International Conference on Computer Vision, Sydney , NSW, pp. 1689-1696, 2013.

[33] **J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu.,** "*Freeform region description with second-order pooling",* 2015, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 37, pp. 1177–1189, 2015

[34] **L. Gong, T.Wang, and F. Liu.,** " *Shape of Gaussians as feature descripors",* 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2366–2371, 2009

[35] **B. Ma, Q. Li, and H. Chang.,** " *Gaussian descriptor based on local features for person re-identification",* 2014 Asian Conference on Computer Vision (ACCV)Workshop, pp. 505-518, 2014

[36] **H. Nakayama, T. Harada, and Y. Kuniyoshi.,** " *Global Gaussian approach for scene categorization using information geometry",* 2010 Conference on Computer Vision and Pattern Recognition, pp. 2336–2343, 2010.

[37] **G. Serra, C. Grana, M. Manfredi, R. Cucchiara.,** " *GOLD: Gaussian of local descriptors for image representation",* 2015, Computer Vision and Image Understanding, Vol. 134, pp. 22-32, 2015

[38] **G. serra, C. Grana, M. Manfredi, and R. cucchiara,** " *Covariance of covariance features for image classification",* 2014 Proceedings of International Conference on Multimedia Retrieval, p. 411, 2014.

[39] **Wang, P. Li and Q.,** " *Local log-Euclidean covariance matrix (L2ECM) for image representation and its applications",* 2012. European Conference on Computer Vision (ECCV),. pp. 469-482, 2012

[40] **X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang.,** " *Hierarchical Gaussianization for image classification,* 2009. proc. IEEE, pp. 1971-1977, 2009.

[41] **K. Q. Weinberger, J. Blitzer, and L. K. Saul., "***Distance metric learning for large margin nearest neighbor classification",* 2009 The Journal of Machine Learning Research, Vol. 10, pp. 207-244, 2009

[42] **M. Guillaumin, J. Verbeek, and C. Schmid.,** " *Is that you? Metric learning approaches for face identification",* 2009. Proc. IEEE Internnational Conference on Computer Vision, Kyoto, pp.498-505, 2009

[43] **Omar Javed, Khurram Shafique, and Mubarak Shah.,** "*Appearance modeling for tracking in multiple non-overlapping cameras",* 2005 proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 26-33, 2005

[44] **Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin.,** " *Person re-identification: What features are important?",* 2012. Proceedings of the European Conference of Computer Vision (ECCV) Workshops, pp. 391-401, 2012

[45] **Yuning Du, Haizhou Ai, and Shihong Lao.,** " *Evaluation of color spaces for person reidentification",* 2012 Proc. 21st International Conference on Pattern Recognition, Tsukuba, pp. 1371-1374, 2012

[46] **E. D. Cheng, M. Piccardi,** " *Track matching over disjoint camera views based on an incremental major color spectrum histogram",* 2005 Proc. IEEE International Conference on Video and Signal Based Surveillance, pp. 147-152, 2005

[47] **S. K. Shah, A. Bedagkar-Gala.,** "*Part based spatio-temporal color appearance model for multi-person re-identification",* 2011 Proc. IEEE International Conference on Computer Vision Workshops, pp. 1721-1728, 2011

[48] **S. K. Shah, A. Bedagkar-Gala**,,"*Partbased spatio-temporal model for multi-person re-identification",* 2012, Pattern Recognition Letters, Vol. 33, Issue 14, pp. 1908-1915, 2012

[49] **Schmid, Cordelia.,** "*Constructing models for content-based image retrieval",* 2001. Proc. IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2, pp. 39-45, 2001

[50] **B. Prosser, W. Zheng, S. Gong, and T. Xiang.,** " *Person re-identification by Support Vector Ranking",* 2010. Proceedings of the British Machine Vision Conference, pp. 1-11, 2010

[51] **Lowe, David G,** *"Distinctive image features from scale-invariant keypoints",* 2004, International Journal of Computer Vision, Vol. 60, pp. 91-110, 2004

[52] **R. Zhao, W. Ouyang and X. Wang.,** "*Unsupervised Salience Learning for Person Re-identification",* 2013. IEEE Conference on Computer Vision and Pattern Recognition, Portland, pp. 3586-3593 2013

[53] **Wen Wang, R. Wang, Z. Huang, S. Shan and X. Chen.,** "*Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets",* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 2048-2057, 2015

[54] **V. Arsigny, P. Fillard, X. Pennec, and N. Ayache,** " *Fast and Simple Computations on Tensors with Log-Euclidean Metrics",* 2005 Proc. international conference on Medical Image Computing and Computer-Assisted Intervention, pp. 115-122, 2005

[55] **Pennec, X., Fillard, P., and Ayache, N.,** "*A Riemannian framework for tensor computing",* 2006, International Journal of Computer Vision, Vol. 66, pp. 41-66, 2006

[56] **Joachims, Thorsten.,** "*Optimizing search engines using clickthrough data",* 2002. Proc. ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 133-142.

[57] **Wei-Shi Zheng, Shaogang Gong, and Tao Xiang.,** "*Person re-identification by probabilistic relative distance comparison",* 2011. Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 649-656, 2011

[58] **Mignon, Alexis.,** "*Pcca: A new approach for distance learning from sparse pairwise constraints",* 2012. Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2666-2672, 2012

[59] **B. Moghaddam, T. Jebara, and A. Pentland.,** " *Bayesian face recognition",* 2000 Pattern Recognition, Vol. 33, pp. 1771-1782.

[60] **T. Hastie, R. Tibshirani, and J. Friedman.,** " *The Elements of Statistical Learning: Data Mining, Inference, and Prediction",* Springer, 2009

[61] **T. Matsukawa, T. Okabe, E. Suzuki and Y. Sato.,** " *Hierarchical Gaussian Descriptor for Person Re-identification",* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 1363-1372, 2016

[62] **N. Otsu, T. Kobayashi,** " *Image feature extraction using gradient local auto-correlations",* 2008. European Conference on Computer Vision (ECCV). pp. 346-358, 2008

[63] **Huang, Z., Wang, R., Shan, S., and Chen, X.,** " *Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning",* 2015, Pattern Recognition, Vol. 48, pp. 3113-3124, 2015

[64] **Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen.,** "*Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification",* 2015 International Conference on Machine Learning (ICML),. pp. 720-729, 2015

[65] **J. Sanchez, F. Perronnin, T. Mensink, and J. J. Verbeek.,** "*Image classification with Fisher vectors Theory and Practice",* 2013, IJCV, Vol. 105, pp. 222-245, 2013

[66] **W. Li, R. Zhao, and X. Wang.,** "*Human reidentification with transferred metric learning",* 2012 Asian Conference on Computer Vision, pp. 31-44, 2012

[67] **C. C. Loy, T. Xiang, and S. Gong.,** "*Time-delayed correlation analysis for multi-camera activity understanding",* 2010 International Journal of Computer Vision, pp. 106-129, 2010

[68] **M. S. Biagio, M. Crocco, M. Cristani, S. Martelli, and V. Murino.,** " *Heterogeneous auto-similarities of characteristics (HASC): exploiting relational information for classification",* 2013 IEEE International Conference on Computer Vision. pp. 809–816, 2013

[69] **F. Xiong, M. Gou, O. camps, and M. Sznaier.,** " *Person re-identification using kerne based metric learning methods",* 2014 Europian conference on computer vision (ECCV), pp. 1-16, 2014

[70] **S. Paisitkriangkrai, C. Shen and A. van den Hengel.,** "*Learning to rank in person re-identification with metric ensembles",* 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846-1855, 2015

[71] **Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li.,** " *Salient Color names for person re-identification",* 2014 European Conference on Computer Vision (ECCV), pp. 536-551, 2014

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