UNIVERSITY OF CALIFORNIA
RIVERSIDE

Active Learning in Multi-Camera Networks, With Applications in Person Re-Identification

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Abir Das

December 2015

Dissertation Committee:

Professor Amit K. Roy-Chowdhury, Chairperson
Professor Anastasios Mourikis
Professor Walid Najjar

The Dissertation of Abir Das is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to acknowledge a number of people who have made me accomplish this journey of PhD. First of all, I thank my advisor Prof. Amit K. Roy-Chowdhury for giving a novish like me the opportunity to work on the utterly interesting topic of person re-identification. Instead of writing a long essay on his amazing capabilities as an advisor I would just say that he is very close to the ideal advisor I could have hoped for. He provided me a lot of help and guidance on my study, research, career and life in general. I consider myself fortunate to be in touch with a person like him.

Next, I would like to thank my committee members Prof. Anastasios Mourikis and Prof. Walid Najjar for their suggestions, comments and insights on how to improve the quality of this dissertation. Research in computer vision is surely a group effort, and this is so very true in our group. It was a privilege working with some of the smart and hard-working individuals who helped in creating a vibrant atmosphere in the lab. My special thanks to Dr. Anirban Chakraborty, Dr. Ramya Malur Srinivasan, Dr. Shu Zhang, Dr. Yingying Zhu, Dr. Nandita Nayak, Dr. Ahmed Tashrif Kamal and Dr. Chong Ding for their continued support and encouragement. They have always been a source of constant support and inspiration for me. I enjoyed interacting with the junior members of the lab. Through the brainstorming sessions with them, I have learnt a lot. In particular I thank Mahmudul Hasan, Jawadul Bappy and Rameswar Panda for their patience with my crazy ideas and for their invaluable advice, friendship, and counsel over the last few years.

I would like to thank my strong network of friends here in Riverside and in India for their emotional support. I'm greatly indebted to my apartment mate Dr. Tusar Tirtha Saha. I have got invaluable guidance in every aspect of life whenever there was need for that.

Last but not the least I owe a lot to my family members. I wish I could imbibe the great quality of taking life as it comes from my elder brother Anjan. He and my sister-in-law were always there with their unconditional support in pursuing my career goals. Words cannot describe the role of my mother Smt. Manju Das and that of my father Sri. Asim Kumar Das. They have always boosted my confidence, corrected me and offered me great moral strength. It wouldn't have been possible for me to make the journey without you Ma (mother) and Baba (father).

Acknowledgment of previously published or submitted materials: The text of this dissertation, in part or in full, is a reprint of the material as it appears in three previously published or submitted papers that I first authored. The co-author Dr. Amit K. Roy-Chowdhury, listed in all these publications, directed and supervised the research which forms the basis for this dissertation. The papers are, as follows.

1. Re-Identification in the Function Space of Feature Warps, in IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015. My co-author Dr. Niki Martinel contributed to the experimentation and analysis.

2. Person Re-identification through Sparse Non-redundant Representative Selection. To be submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence. My co-author Rameswar Panda contributed to the writing.

3. Active Image Pair Selection for Continuous Person Re-identification, in IEEE International Conference on Image Processing, 2015. My co-author Rameswar Panda contributed to the writing.

To my parents.

ABSTRACT OF THE DISSERTATION

Active Learning in Multi-Camera Networks, With Applications in Person Re-Identification

by

Abir Das

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2015
Professor Amit K. Roy-Chowdhury, Chairperson

With the proliferation of cheap visual sensors, camera networks are everywhere. The ubiquitous presence of cameras opens the door for cutting edge research in processing and analysis of the huge video data generated by such large-scale camera networks. Re-identification of persons coming in and out of the cameras is an important task. This has remained a challenge to the community for a variety of reasons such as change of scale, illumination, resolution *etc.* between cameras. All these leads to transformation of features between cameras which makes re-identification a challenging task. The first question that is addressed in this dissertation is - *Can we model the way features get transformed between cameras and use it to our advantage to re-identify persons between cameras with non-overlapping views?* The similarity between the feature histograms and time series data motivated us to apply the principle of Dynamic Time Warping to study the transformation of features by warping the feature space. After capturing the feature warps, describing the transformation of features the variabilities of the warp functions were modeled as a function space of these feature warps. The function space not only allowed us to model feasible transformation between pairs of instances of the same target, but also to separate them from the infeasible transformations between instances of different targets. A supervised

training phase is employed to learn a discriminating surface between these two classes in the function space.

However, it is unlikely that supervised methods alone will be enough to deal with the volume and variety of data in such scenarios. The performance is dependent on tediously labeling the training data. Also supervised person re-identification strategies are static in the sense that these are unable to adapt to the changing dynamics of continuous streaming data. Active participation of human expert is necessary in such scenario. The human labor is reduced if the human is involved for the most difficult cases and if it can be made sure that the human expert is not asked to do the same job repetitively. So the question we addressed is the following. *Is it possible to identify a manageable set of informative, but non-redundant, samples for labeling by a human expert? Moreover, is it possible to select these examples progressively in an online setting where all the training data may not be available a priori?* The dissertation explored a convex optimization based iterative framework that progressively and judiciously chooses a sparse but informative set of samples for labeling, with minimal overlap with previously labeled images. The third part of the dissertation also addresses the same basic question from a different perspective where the human effort is reduced in two ways - by changing the questions asked to the human annotator to binary yes-no type instead of multiple choice and by incorporating the domain knowledge from the human how a human expert discriminates between persons. The two objectives are fulfilled by employing a 'value of information' based active learning strategy and mid level semantic 'attributes' respectively. Via extensive experimentation with different scenarios, we validate our approach and demonstrate that our framework achieves superior performance with significantly less amount of manual labor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ubiquitous reach of cameras in almost every aspect of human life have resulted in huge amount of visual data. Nearly 112 million users of the photo sharing engine 'flickr' share a million images daily on average, while 300 hours of video are getting uploaded every minute in 'youtube'. Such a flurry of visual data has ushered in a new era in computer vision research. Visual surveillance has been one of the most active application areas emerging out of the deployment of camera networks. Networks of vision sensors are deployed in many settings, ranging from security needs to disaster response to environmental monitoring to monitoring patients, elderly and children [56, 82]. Both cheap price of visual sensors and processors as well as growing need of public safety and security are drivers of the growing interest in video data analysis. This raises the need for automated methods able to extract, and access high-level semantic information carried by the extremely high volume of recorded video data.

Due to the easy availability of a multitude of visual sensors, a new domain of computer vision research has evolved which focuses on monitoring crowded and busy scenes with a network of cameras rather than processing only one video stream with a single camera view. Even though the sensing devices are becoming cheaper, covering a wide area 'fully'

by deploying a large number of cameras is still not feasible due to the amount of human supervision, privacy concerns, and maintenance costs involved. As a result, only a small part of the whole area is covered by a number of cameras with non-overlapping fields-of-view (FoVs). The non-overlapping camera FoVs leave *"blind gaps"* which are critical in the sense that no information can be obtained from these areas. As a result of losing a person when he/she leaves a camera FoV, it is extremely challenging to re-associate the same person at a different location and time among multiple persons. This inter-camera person association problem is known as the person re-identification problem. Fig. (1.1) shows an example scenario of person re-identification considering a 3 camera network. The camera FOVs are non-overlapping. In between appearing into two camera FOVs, different persons can be in any of the blind gaps. A person re-identification algorithm takes two images from two non-overlapping cameras and provides a decision whether those two images are of the same person or not. The decision of match or non-match is taken based on the re-identification algorithm giving a probability of match or mismatch between the two images.

1.1 Challenges

In spite of a surge of effort put in by the research community in recent years, re-identification has remained quite an open issue due to a number of hard challenges. Firstly, footages are recorded in an uncontrolled environment by cameras with large FoVs, generating low resolution images of the targets. This makes the acquisition of discriminating biometric features (*e.g.*, face and gait features) hard as well as unreliable. Due to the poor quality of the acquired biometric features, methods relying on such features perform unsatisfactorily. As a result, visual appearance features are, still, the first choice in re-identification problems. As a target's appearance often undergoes large variations across

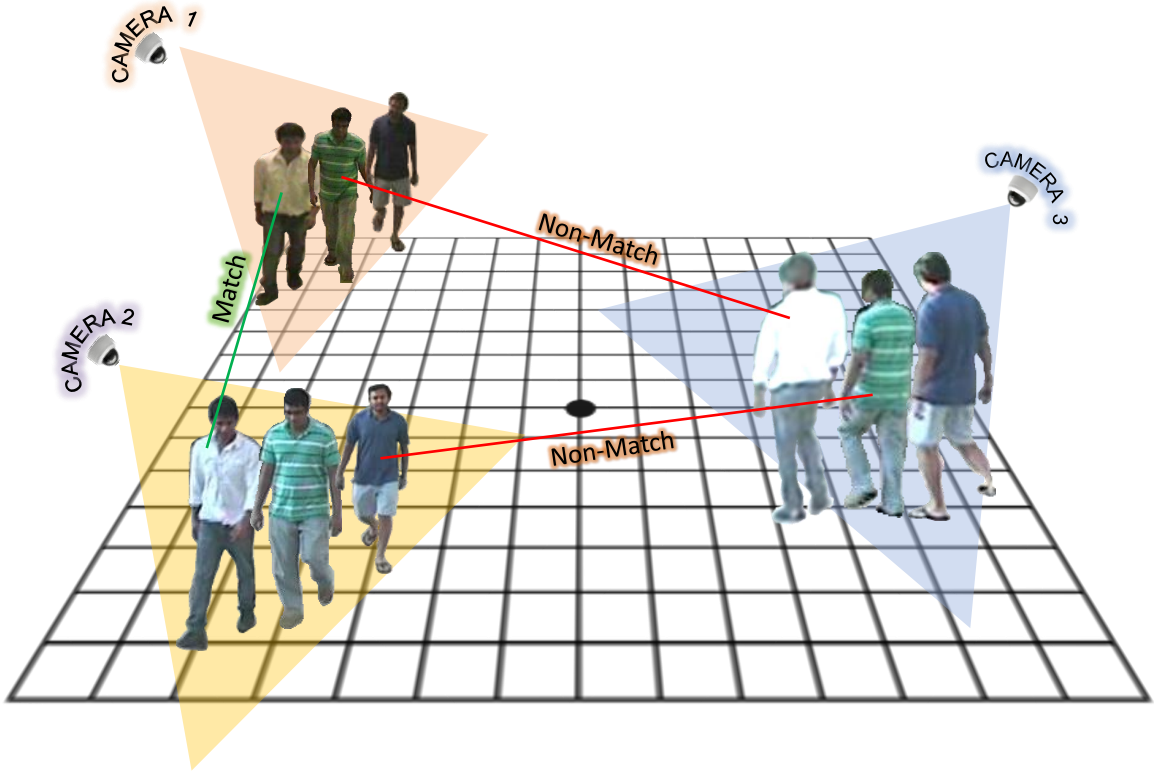


Figure 1.1: Person re-identification. Person re-identification algorithms assume non-overlapping camera FOVs and movement of persons from one camera FOV to another via blind gaps. A person re-identification algorithm takes two images from two non-overlapping cameras and provides a decision whether those two images are of the same person or not based on a probability score of match or non-match between the two images.

non-overlapping camera views due to significant changes in viewing angle, lighting, background clutter, and occlusion, the appearance features for the target can be very different from camera to camera. This is especially true in case of person re-identification due to the non-rigid shape of the human body. An example of such a scenario is shown in Fig. 1.2. Three frames of the same person acquired by three non-overlapping cameras are presented together with the color histogram (hue, saturation and value) features. As shown, such features are significantly different for the same target viewed in different cameras making the re-identification problem very challenging.

In such a scenario, information about time of travel between different camera FOVs can be handy. For a smaller camera network the time information can be used effectively

Figure 1.2: Three images of the same person in three non-overlapping cameras from the RAiD dataset [21]. Below each image, HSV features are shown as 3 different histograms. Brown denotes the hue, green denotes saturation and the sky-blue denotes the value histograms respectively. The inconsistency of the histogram shape does not allow them to be used as unique features for re-identification.

to find the camera topology [104] and the learned topology can be employed to discard candidate images for which the time informations do not conform to the general travel time between the corresponding camera FOVs. However, for large camera network covering wide area, there can be many paths of travel from one camera FOV to another. This results in multimodal distributions of travel time between cameras which can be an unreliable cues due to the large number of possible paths between cameras.

Traditional methods of person re-identification addressing these challenges, involve an intensive supervised training phase [7, 17, 26, 45, 112, 116] where it is assumed that all the training examples are labeled. This approaches try to capture large appearance variations of different persons across cameras by labeling as many images of them as possible. Considering the time, labor and human expertise involved in labeling the training data manually, person

re-identification for a large number of persons, often, suffers from the curse of scalability when using traditional but otherwise tested approaches. Some recent semisupervised or unsupervised methods [55, 73, 118] tried to reduce human labor of labeling or annotating training data by exploring saliency information or weighted features towards re-identifying people across cameras. Though for small scale systems the performance is good, the lack of the study of scalability of performance with number of persons or sensors or number of detections per person calls for involving the best working system - *human*, efficiently. Though human help is effective for large scale systems, it is costly.

Active learning [102] based systems have been used to involve human in the loop where labels for only those examples are sought from the human which the system finds difficult. Involving human only for difficult examples rather than for samples chosen randomly from a set or for the whole set, reduces the human effort as only a small part of the whole data is annotated. Active learning based methods [15, 30, 50], though, have been studied for various problems with data coming from single source, it is not trivial for a person re-identification scenario where multi-sensor data is considered. Since multiple sensors can capture the images of the same person, many images can be redundant. So in presence of multiple sensors as in person re-identification scenarios, judicious pruning of redundant examples will be effective in reducing human labor of annotation as this makes sure that the human annotator does not label the same person repetitively. It is a natural challenge to select a few informative samples yet cover as much appearance variation as possible across multiple cameras in such a scenario.

Reduction of redundancy is of utmost importance especially when all the data may not be available at the very outset. An example scenario can be where the input is images of detected persons from streaming videos. A static pre-trained model can not adapt to the changing dynamics of the incoming data. In such a case choosing difficult

examples progressively for annotation, as and when new data arrives, is necessary. Such an iterative framework will not be efficient if images of the same person is chosen repetitively for annotation. So the challenge will be to select a manageable set of training images for annotation from multi-sensor data in an online person re-identification scenario.

Traditional active learning settings ask for labels from the human expert. While label information is sufficient to train the model, an effective way to reduce human effort is to teach the system how a human discriminates between persons. Though such incorporation of the domain knowledge of the expert in the process can help in reducing the subsequent effort in labeling, it requires to resort to a mid level language which is understood by both man and the machine. A recent line of work [32, 87, 88] draws inspiration from the way human experts simplify the task of discrimination by using mid level semantic features, called *attributes*. Unlike low level features (*e.g.*, HOG [20], SIFT [74] *etc.*) which are machine detectable or high level concepts like person identities which are human understandable, attributes are both machine detectable (*i.e.*, machines can be trained to detect attributes) and human understandable. Soft-biometric features like 'having long hair' or appearance features like 'wearing green colored shirt or not' helps a human expert to discriminate between several persons and these can be used as attributes.

In person re-identification [64, 65, 105], attributes have only been used in the supervised setting as a replacement for the low level features. However, for an online re-identification system starting with absolutely no attribute information, the challenge is to incorporate the domain knowledge from human in terms of attributes and simplify the task of discriminating between persons.

1.2 Contributions of the Thesis

The objective of this work is to design algorithms that can address change of features between cameras and involve human efficiently in order to achieve robust multi-camera person re-identification. The study has been carried out in three parts.

First, we explore the way feature gets transformed between cameras as a result of variations in viewing angle, lighting, scale resolution *etc.* and use the learned knowledge to find out whether two persons from two different cameras are same or not. To capture the transformation, we made use of the similarity of transformation of features between cameras and transformation of time sequences especially studied in the speech and audio signal processing community. Dynamic Time Warping or DTW [11, 83] has been widely used to study the nature of transformation from one time sequence to another by morphing or warping the time axis. Time seqs are functions of time and feature histograms are functions of bin numbers. As a result, using a similar principle to DTW, we model the transformation of features by non-linearly warping the feature space to get the "*warp functions*". The warp functions between two instances of the same target from two non-overlapping cameras form the set of feasible warp functions while those between instances of different targets form the set of infeasible warp functions. We build upon the observation that feature transformations between cameras lie in a nonlinear function space of all possible feature transformations. The space consisting of all the feasible and infeasible warp functions is the warp function space (WFS). We propose to learn a discriminating surface separating these two sets of warp functions in the WFS and to re-identify persons by classifying a test warp function as feasible or infeasible. Towards this objective, a Random Forest (RF) classifier is employed which effectively chooses the warp function components according to their importance in separating the feasible and the infeasible warp functions in the WFS. Extensive experiments

on five datasets are carried out to show the superior performance of the proposed approach over state-of-the-art person re-identification methods. In addition, it has been shown that the proposed method reaches the best average performance over multiple combinations of the datasets, thus, showing that the method is not designed only to address a specific challenge posed by a particular dataset.

Next, we extend person re-identification to 'continuous person re-identification' which would not rely on a statically learned model on tediously labeled training data. Traditional supervised re-identification methods are static and will not be suitable when new data arrives continuously or when all the data is not available for labeling beforehand. As a result we involve human in the loop for continuous and online person re-identification. But involving human is costly. In this second work we involve human in an efficient manner for continuous person re-identification by selecting a small set of informative and non-redundant samples for annotation by the human experts. For large multi-sensor data as typically encountered in person re-identification, labeling a lot of samples does not always mean more information, as redundant images are labeled several times. In this work, we propose a convex optimization based iterative framework that progressively and judiciously chooses a sparse but informative set of samples for labeling, with minimal overlap with previously labeled images. Another issue in such an online iterative framework is the quick adaptability of the system to the changing dynamics of the incoming data. Though discriminative classifiers (*e.g.*, SVM or random forest) have shown good classification performance, the generally exponential increase of training time with the number of training samples for them is a hindrance to the scalable solution of the problem. These classifiers have to be retrained from scratch after each batch of representative selection and annotation in such repetitive active learning strategy. Motivated by the success of Sparse Reconstruction Based Classifiers (SRC) [24, 113], we used a structure preserving SRC to reduce the training burden

typically seen in these discriminative classifiers. The two stage framework not only helps in reducing the labeling effort but also can handle situations when new unlabeled data arrives continuously. This is due to the fact that online update of the classifier involves only the incorporation of new labeled data rather than any expensive training phase. Using three benchmark datasets, we validate our approach and demonstrate that our framework achieves superior performance with significantly less amount of manual labeling.

Continuing on involving human efficiently, the third work addresses continuous person re-identification from a different perspective. In particular, the human effort is more when it has to answer a multiple choice question compared to a binary yes-no type question. Inspired by the 'value of information' active learning framework [50], we propose a continuous learning person re-identification system which provides the human expert one unlabeled image and another labeled image and asks the human if the pair matches or not. An information theoretic criterion judiciously chooses the image pair so that such response from the expert will facilitate the system to gain most information. The human in the loop not only provides labels to the incoming images but also improves the learned model by providing most appropriate attribute based explanations. These attribute based explanations are used to learn attribute predictors along the way, as opposed to using a predetermined set of attributes. This leads to a framework for selecting an optimal order of images for labeling, as well as an effective set of attributes. The overall effect of such a strategy is to limit the labeling effort of the human. Using two benchmark datasets, we validate our approach, in terms of accuracy and manual labeling effort, and compare against state-of-the-art methods.

1.3 Related Works

We provide a review of related work on person re-identification, active learning and attribute based learning.

1.3.1 Person Re-identification

In the last few years the problem of re-identifying persons across multiple non-overlapping cameras has received increasing attention. The community has commonly adopted three different kind of approaches: i) discriminative signatures based methods, ii) metric learning based methods, iii) feature transformation learning based methods. A multidimensional taxonomy and categorization of the person re-identification algorithms can be obtained in the review paper [109]. In the rest of the subsection we do a thorough review of the existing re-identification works.

Discriminative signature based methods [7, 17, 72, 80] use multiple standard features *e.g.*, color, shape, texture *etc.* or specially learned features like Biologically Inspired Features (BIF) [77], covariance descriptors [6], shape descriptors of color distributions in log-chromaticity space [59], deep features [71] *etc.* to compute discriminative signatures for each person using multiple images. Some recent methods have shown that representing the query images based on reference datasets [3, 114] can be used to boost the re-identification performance.

According to [26], in a metric learning framework a set of training data is used to learn a non-Euclidean metric which minimizes the distance between features of pairs of true matches while maximizing the same between pairs of wrong matches. Works trying to improve the metric learning performance by excluding well separable examples and solving an eigenvalue problem [44], by giving less importance to unfamiliar matches in a large margin

nearest neighbor framework [26], by learning multiple metrics specific to different candidate sets in a transfer learning set up [70] or by exploiting sparse pairwise similarity/dissimilarity constraints [81] have shown remarkable re-identification performance. Metric learning based person re-identification has also been formulated as a local distance comparison problem on energy-based loss functions [57, 117] or Local Fisher Discriminant Analysis [91]. To reduce the computational costs, a relaxation of the positivity constraint of the Mahalanobis metric has been proposed [45]. The interested readers are directed to two survey papers [10, 115] for a detailed description of metric learning approaches.

A similar approach to metric learning is dissimilarity measure learning [90] which has been used successfully in person re-identification [98, 99]. These methods create a set of dissimilarity descriptors based on a set of visual prototypes obtained by unsupervised clustering. Person re-identification is, then, formulated as a supervised classification problem with the learned dissimilarity descriptors as features. Recently, both discriminative feature learning and metric learning have been treated as a joint learning problem employing deep convolutional architecture providing competitive performance [2]. However, similar to many other deep architecture, generating huge amount of labeled training data is an issue which has been addressed in the third part of the thesis.

A third class of works tried to explore transformation of features between cameras by learning brightness transfer function [48, 93] between appearance features. Later an incremental learning framework modeling linear color variations [40] between cameras were used to match the targets. Both [40] and [48] learned space-time probabilities of moving targets between cameras and used them as cues for association. However, transition time information may be unreliable if camera FoVs are significantly non-overlapping. Efforts of improving the BTF resulted in a BTF modeling the effects of illumination changes over time [103], a sparse color information preserving Cumulative BTF [94], or a Weighted BTF

designed to assign unequal weights to observations based on how close they are to test observations [22]. In [5] the re-identification problem was posed as a classification problem in the feature space formed of concatenated features of persons viewed in two different cameras.

1.3.2 Active Learning

In an effort to bypass tedious labeling of training data there has been recent interest in 'active learning' [50, 110] to intelligently select unlabeled examples for the experts to label in an interactive manner. The specific application areas in computer vision include, but not limited to, tracking [111], scene classification [50, 110], semantic segmentation [108] and video annotation [52]. Queries are selected for labeling such that enough training samples are procured in minimal effort. This can be achieved by choosing one sample at a time by maximizing the value of information [50], reducing the expected error [4], or minimizing the resultant entropy of the system [13] or maximizing both informativeness and representativeness for active sample selection [47] prior to retraining a classifier. On the other hand there have been recent approaches [15, 30] where batches of unlabeled data are selected by exploiting classifier feedback to maximize informativeness and sample diversity. For a detailed discussion on active learning literature, the interested readers are directed to the excellent article by Settles [102].

1.3.3 Attribute based Learning

These mid-level semantic concepts have seen a surge of applications in many areas of computer vision. Attributes are describable aspects of information such as a facial expression, age, gender, pose or could be any other side information such as spectacles, beard, facial scar, etc. Generating descriptions of unfamiliar objects have facilitated zero

shot learning and detection, simply by specifying the attributes of the object [32, 62]. In one of the early works [58], the authors proposed attribute classifiers that are basically binary classifiers trained to recognize the presence or absence of attributes. Attributes work as an excellent communication tool between human and machines to facilitate boosted learning experience [61, 89]. Inspired by the success of attribute centric approaches, a recent line of work have studied the use of attributes in re-identification [64, 65, 73, 105]. The principle has been to use manual or crowdsourced annotations to train models and then to use the model generated attributes for recognition. While these works use pre-annotated data with a predefined vocabulary of such attributes, our proposed work uses a set of useful attributes with active labeling from the human in the loop.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter 2 addresses the problem of multi-camera target re-identification by learning the way different features get transformed between two cameras. Given a pair of feature vectors we show that we can learn the decision surface best separating the feasible and infeasible set of such feature transformations. The target re-identification problem is formulated as determining whether two images from different cameras belong to the feasible or infeasible regions of the space of all feature transformations. In Chapter 3, we extend person re-identification as a continuous learning system involving human in the loop with an eye to reduce human labor by progressively and judiciously choosing a sparse but non-redundant set of samples for labeling. A sparse representation based classifier is employed to facilitate online updation of the model without having to be limited by the knowledge of the number of classes from the start or without having to retrain from scratch after each batch of data arrives for training. Chapter

4 proposes an information theory based strategy to reduce human labor of annotation using a value of information based active learning strategy with attribute feedback. We provide a summary of the thesis and highlight directions for future work in Chapter 5.

Chapter 2

Re-Identification in the Function Space of Feature Warps

In this chapter, we will discuss why feature transformation makes the problem of re-identifying persons across non-overlapping cameras hard. Next, we will provide a brief review of literature on the study of transformation of features and its use in person re-identification. We will then describe the overall framework involving a) Feature extraction b) Computation of feature transformation based on the principles of DTW to get feasible and infeasible warp functions c) Training a binary classifier with the computed warp functions and d) Re-identification posed as classifying a test warp function to be feasible or infeasible. Next we provide results of extensive experiments on five datasets and discuss the significance of the results.

2.1 Introduction

The last few years have seen a lot of research effort put in re-identifying people across multiple non-overlapping cameras. In spite of that, re-identification, still, is a hard

challenge due to a number of reasons. One of them is the low quality of surveillance videos typically encountered in real life re-identification scenario. Due to the poor quality of the acquired videos acquisition of discriminating biometric features (e.g. face and gait features) hard as well as unreliable. As a result, visual appearance features are, still, the first choice in re-identification problems. As a target's appearance often undergoes large variations across non-overlapping camera views due to significant changes in viewing angle, lighting, background clutter, and occlusion, the appearance features for the target can be very different from camera to camera. An example of such a scenario is shown in Fig. 1.2 in chapter 1 where it was shown how simple appearance features get changed across cameras.

The computer vision community has tried to address the re-identification problem by designing discriminative signatures for each target or by finding a non-Euclidean metric which minimizes distance between the features of the same target across cameras. These methods rely on the fact that the individual signatures vary a little from camera to camera. Such methods, while efficient and effective to re-identify persons viewed in different poses, result in a significant loss of performance when strong illumination and color changes occur between different cameras. As a result of these changes, features describing the same person get transformed between cameras. Thus an important aspect of the problem is to understand how features get transformed across cameras. Fig. 2.1 shows an example where a person goes from a brightly illuminated space (Fig. 2.1(a)) to a dark place (Fig. 2.1(b)). This large change of illumination is also depicted by the shift of the distribution of pixels from the higher end values towards the lower end values in the corresponding grayscale histograms (shown alongside the two images). This change in the shape of the distribution can be captured by studying the histogram warp. We use the principles of Dynamic Time Warping (DTW) for this purpose. DTW [11, 83] is a dynamic programming algorithm that

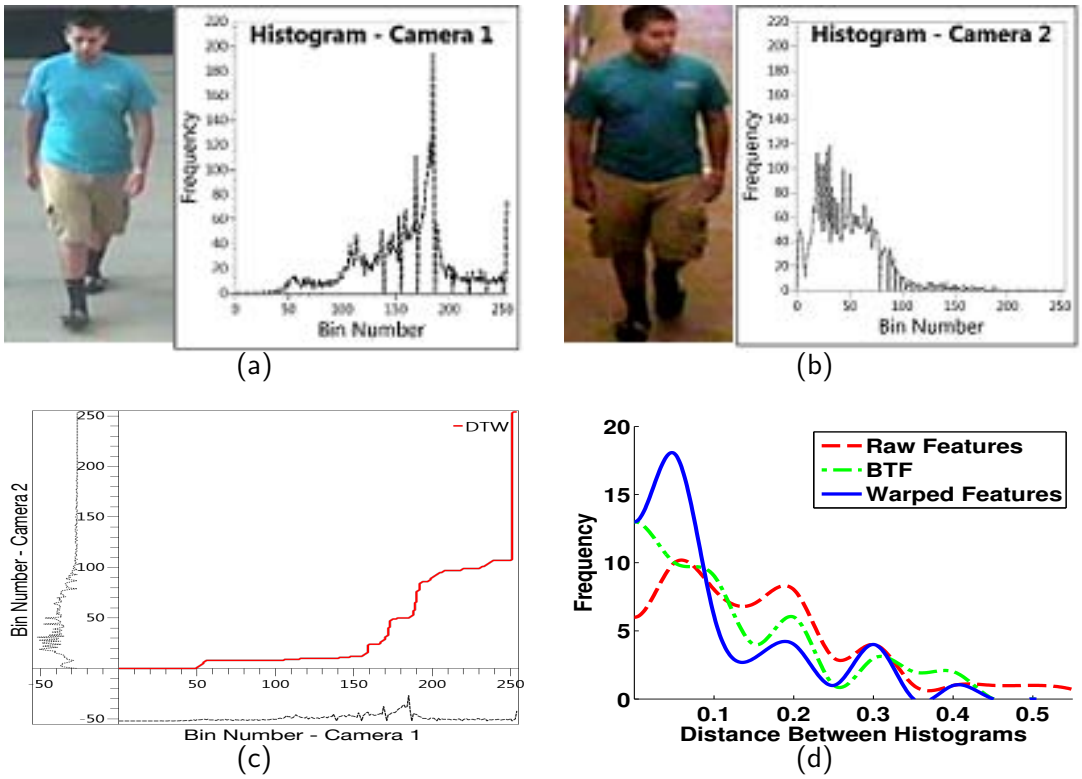


Figure 2.1: Using the principle of DTW to capture the transformation of features as a person goes from a brightly illuminated space to a dark place. (a) and (b) show the images of a person along with its value histogram plots at a brightly illuminated and a dark place respectively. (c) shows the warp function which maps the bin numbers of the color histogram in (a) to the bin numbers of the color histogram in (b). The initial flatness and latter steepness of the warp function captures the transformation of features resulting from the change in illumination. Fig. (d) shows the distribution of the Bhattacharyya distances between the transformed and actual grayscale histograms using BTF [48] (in green) and warp functions (in blue) computed for all the 50 persons in the CAVIAR4REID dataset. Concentration of more persons with smaller distances using warp function can be readily seen. The distribution of the distances computed between the raw value histograms is also shown for comparison (in red).

optimizes the alignment of two time series by non-linearly warping the series so that the sum of the point-to-point distances is minimized. Time sequences are functions of time while color histograms are functions of the bin numbers. So the same principle can be used to study the warping of the bin number axis causing the change in the shape of the distributions. Fig. 2.1(c) shows such a *warp function* which captures the feature transformation by mapping the bin numbers of the color histogram in Fig. 2.1(a) (shown as the horizontal axis) to the bin numbers of the color histogram in Fig. 2.1(b) (shown as the

vertical axis). The initial flatness and latter steepness of the warp function characterize the shift of the concentration of the pixels from the higher to the lower end of the color histogram. Fig. 2.1(d) shows a comparative performance of the use of warp function and an widely used feature transformation method, the brightness transfer function (BTF) [48] on capturing the feature transformation. Value histogram features of images from one camera in CAVIER4RAID [17] dataset was transformed to features from another camera using warp functions and BTFs. Bhattacharyya distances between the transformed feature and the original feature in the second camera are computed for both the feature transformation methods. As shown by Fig. 2.1(d), the distribution of the number of people for which the distance is smaller is more when warp function is used than when BTF is used to transform the feature from one camera to other.

The existing studies exploiting feature transformation, have tried to learn linear [40] and nonlinear transformation functions [93, 48] between appearance features among pairs of cameras. These approaches, however, use the learned transformation function to project the features from one camera to the feature space of the other camera. In a re-identification scenario this may not always be feasible since the mapping may not be unique and it may vary from frame to frame depending on a large number of camera parameters (e.g. illumination, scene geometry, exposure time, focal length, and aperture size). In this work, we build upon a detailed understanding of the transformation of features captured by warp functions computed based on the principles of DTW. Considering two non-overlapping cameras, a pair of images of the same target is denoted as a feasible pair, while a pair of images between two different targets is denoted as an infeasible pair. The corresponding warp functions describing the transformation of features are denoted as *feasible* (positive) and *infeasible* (negative) warp functions respectively. The set of infeasible warp functions vary widely as in this set the warps are computed for image pairs consisting of different

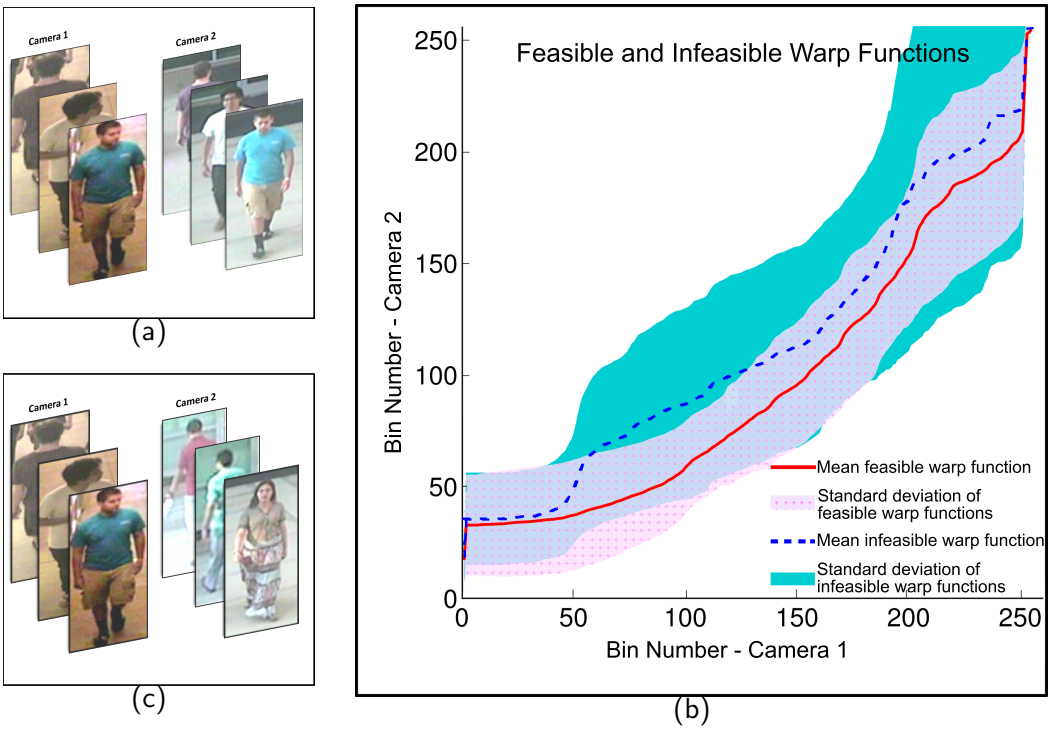


Figure 2.2: Feasible and infeasible warp functions in the WFS. (a) and (c) show example images of the feasible and infeasible pairs respectively taken from an outdoor and an indoor camera of the RAiD [21] dataset. Fig. (b) shows the mean of the feasible (in bold line) and infeasible warp functions (in dashed line) between the grayscale histograms of the torso of the feasible and infeasible pairs. 100 randomly chosen examples of feasible and infeasible warp functions are averaged to get the mean warp functions. The shaded areas show the corresponding spread of the variances (as $\pm$ standard deviation value). This figure shows that feasible and infeasible warp functions for this simple feature (grayscale histogram) can be discriminative and can be used for re-identification.

persons. Even within the set of feasible warp functions, the transformations are not unique when computed for different feasible pairs. For each of the two sets, the feature transformations may not be well represented by a single warp function in presence of such variabilities. So, we propose to model the function space capturing all the feasible and infeasible warp functions between pairs of cameras, termed as the feature *warp function space* (WFS). The WFS not only allows us to model feasible transformation between pairs of instances of the same target, but also to separate them from the infeasible transformations between instances of different targets. This enables us to address the re-identification problem as a binary classification problem by discriminating in the WFS.

Fig. 2.2 shows a visual proof of the discriminating power of the feasible and infeasible warp functions. For convenience of visualization, we resorted to a low dimensional WFS by computing the warp functions between the grayscale histograms of the images. Fig. 2.2(a) and (c) respectively show some example feasible and infeasible image pairs from the RAiD [21] dataset corresponding to camera 1 and 3. Since, in general, people wear different colored clothes for torso and legs the warp functions for the two bodyparts are computed separately. For visualization convenience we show the warp functions for torso only. Fig. 2.2(b) shows the mean feasible (in bold line) and infeasible warp functions (in dashed line) between the grayscale histograms of 100 randomly chosen feasible and infeasible pairs of images respectively. The shaded areas show the corresponding spread of the variances (as $\pm$ standard deviation value). This shows that both the mean warp function and the spread of variance are different for feasible and infeasible warp functions even for this simple feature (grayscale histogram). The proposed work explores this discriminating power of the feasible and infeasible warp functions in the WFS for person re-identification. Since, most of the benchmark datasets include changes of scale and viewpoint in addition to illumination, it may not always be possible to discriminate well enough using such a simple feature representation. So, we computed the warp functions between other dense color and texture features in the actual experimentations to deal with these challenges. Discrimination between the two classes of warp functions are further enhanced in a classification framework which finds a complex discriminating surface in a higher dimensional WFS consisting of the warp functions computed between these features. Details of the feature extraction and computation of warp functions can be obtained in Section 2.4.

To summarize, the contributions of the proposed approach to the problem of person re-identification are the followings. To capture the feature transformation we propose to compute a nonlinear mapping (warp function) that minimizes a cost defined as the mismatch

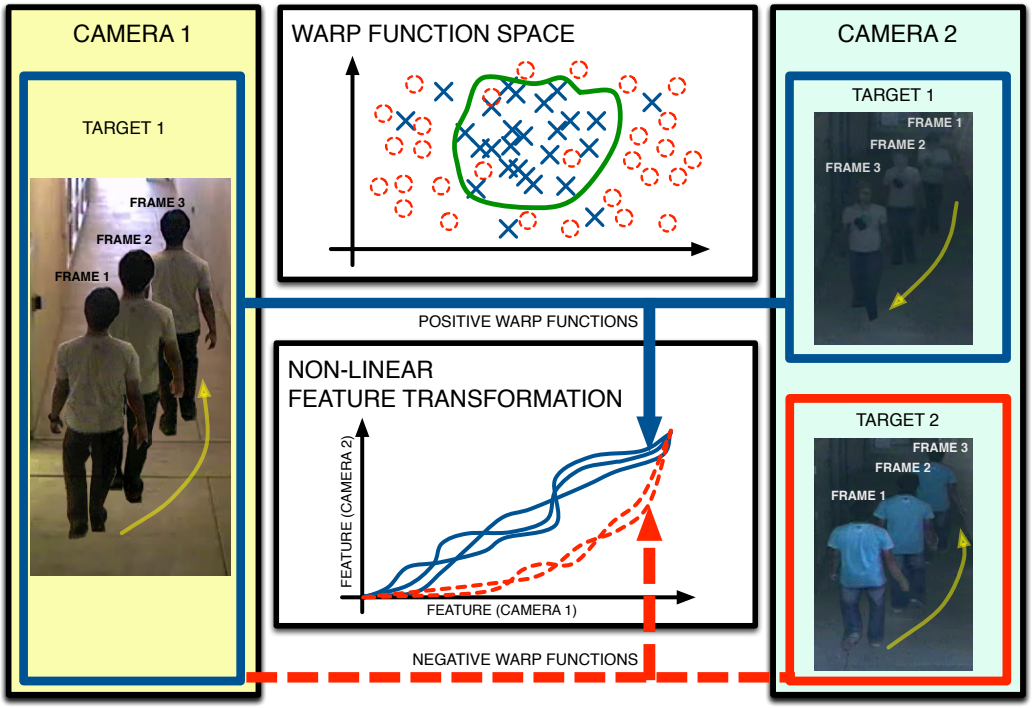


Figure 2.3: Re-identification by discriminating in the warp function space. The warp functions computed between features extracted from images of the same target (*i.e.*, positive warp functions) are shown in solid blue. The warp functions computed between features extracted from different targets (*i.e.*, negative warp functions) are shown in dashed red. A nonlinear decision surface (shown in green) is learned to separate the two regions.

between histogram features. A WFS composed of the collection of feasible and infeasible warp functions is built. We also propose to learn a discriminating surface between the sets of feasible and infeasible warps in the WFS using a random forest of decision trees. The re-identification problem is addressed by mapping a test warp function onto the WFS and classifying it as belonging to either the set of feasible or infeasible warp functions (see Fig. 2.3).

We compare the performance of our approach to state-of-the-art person re-identification methods using five publicly available benchmark datasets. The datasets are chosen with a particular focus on large illumination variation between cameras. Since we learn the space of feature transformations, our results significantly outperform others when applied to datasets with large appearance variations between the cameras, such as RAiD [21] and

WARD [80]. Also, we demonstrate that our method is not tuned to any specific dataset. Our average performance on different combinations of multiple datasets is higher than other state-of-the-art methods.

The rest of the chapter is organized as follows. Section 2.2 gives a brief description of the state-of-the-art approaches in person re-identification. An overview of the proposed approach is given in Section 2.3. The details about the re-identification approach, as feature extraction, warping and WFS are described in Section 2.4. Experimental results and comparisons with state-of-the-art methods are shown in Section 2.5. Finally, conclusions are drawn in Section 2.6.

2.2 Previous works in Person Re-Identification

The person re-identification algorithms can be broadly categorized into three different kinds. They are i) discriminative signatures based methods, where one tries to find camera invariant features that does not vary much between cameras. In this approach the features are complex but the distance calculating functions are simple *e.g.*, Euclidean or Bhattacharyya distances. ii) metric learning based methods, where one takes a complimentary approach. Here the features are kept simple but the metric which calculates the distance between two images using these features are specially designed. The metrics learned in this way aim to minimize the distance between pairs of true matches at the same time try to maximize the distance between two images which are not of the same person. iii) feature transformation learning based methods, which is completely different from the last two. These approaches are based on the fact that features are going to change between cameras and they try to learn the way features get transformed and use this knowledge to re-identify people. Section 1.3 in chapter 1 has a detailed discussion on some of the

state-of-the-art methods in each category. Here we will discuss some of the close works with the proposed method and address the differences with them.

In one of the early works [93] studying the transformation of features, a BTF between appearance features was computed by finding the optimal path in the feature correlation matrix. Later, a learned subspace of the computed BTFs [48] and a incremental learning framework modeling linear color variations [40] between cameras were used to match the targets. Both [40] and [48] learned space-time probabilities of moving targets between cameras and used them as cues for association. However, transition time information may be unreliable if camera FoVs are significantly non-overlapping.

A basic difference of the metric learning or dissimilarity measure based methods with our approach is that these methods do not take into account the transformation of features which is especially useful when there is a significant but consistent change of appearance of the individuals between cameras. Also the methods based on person specific signature, dissimilarity measure and metric learning have to either rely on the assumption that all the persons are seen during the training phase or carefully choose threshold value separating the new persons from the matches with existing persons. Since we are exploiting transformation of features between cameras and it is independent of the specific persons, the proposed method is more general in this sense.

In this work we focus specifically on the issue of how features are transformed between views and learn a model of these transformation functions. We pose the re-identification problem as computing these nonlinear warp functions between features and learning a function space which models the feasible and the infeasible warp functions.

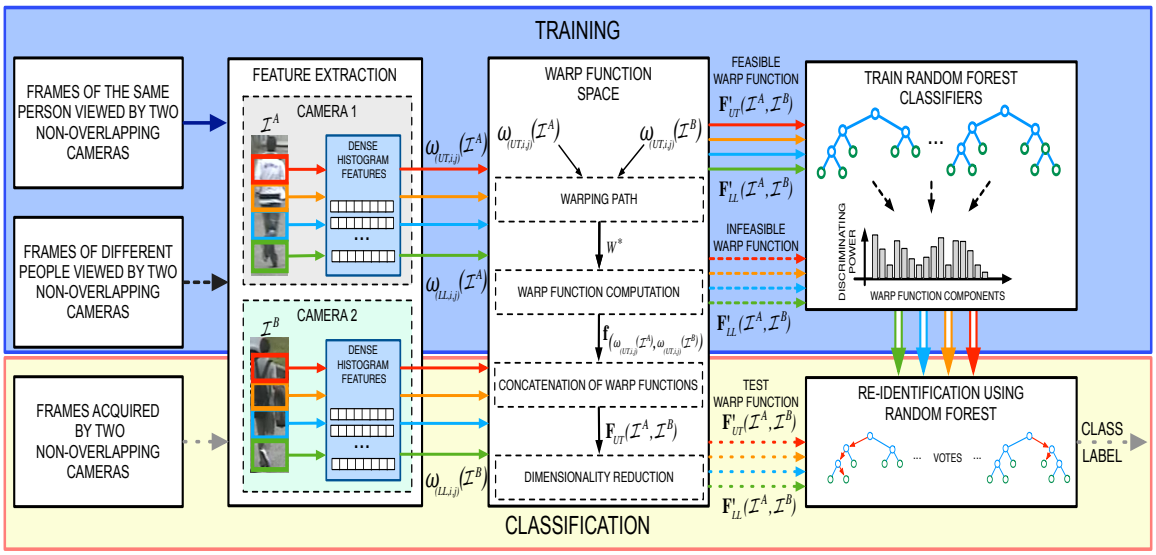


Figure 2.4: System Overview. The feature extraction module takes raw video frames and extracts dense color and texture features from each of the four detected body parts. These are input to the warp function space module that computes the warp function between each of them and reduces the dimensionality of the warp function space. A random forest classifier is trained to discriminate between the feasible and the infeasible warp functions in the WFS. The trained classifier is used to classify the test warp functions.

2.3 Overview of proposed approach

The overall scheme of the proposed person re-identification process is shown in Fig. 2.4. Given the frames from two cameras we learn a discriminative model in the WFS to get the probability of a sample feature warp function coming from the same person or not.

Towards this objective, we first extract features from the person images. The feature extraction module performs the following tasks: a) splitting the image of the detected persons into four main body parts, and b) extracting dense color and texture features from the detected body parts.

For each extracted feature, vector valued warp functions are computed by the warp function space module. All the warp functions (corresponding to different features) are concatenated to form a high dimensional warp function for each image pair. The

warp function between the same target in different cameras is denoted as a feasible or positive warp function while the warp function between two different targets is denoted as an infeasible or a negative warp function. The set of all feasible and infeasible warp functions forms the WFS. The dimensionality of the WFS is reduced using Principal Component Analysis (PCA) [46].

Given the WFS, a decision surface discriminating the two sets of warp functions is learned using a Random Forest (RF) [14] of bagged decision trees. Every component of the warp functions may not be discriminating enough between the two classes of transformations (feasible/infeasible). The decision trees select the subset of warp function components according to their importance and maximize the discrimination between the feasible and infeasible warp functions in the WFS.

For classification, features are extracted from test image pairs and input to the WFS module to compute the warp functions. Finally, the RF classifies the test warp functions in the WFS as feasible or infeasible.

2.4 Methodology

In this section we describe the different modules of the proposed approach in details.

2.4.1 Feature extraction

State-of-the-art methods for person re-identification have successfully explored different appearance features [72]. While existing feature transformation based methods are designed for color features, our framework can be used to study the nature of the trans-

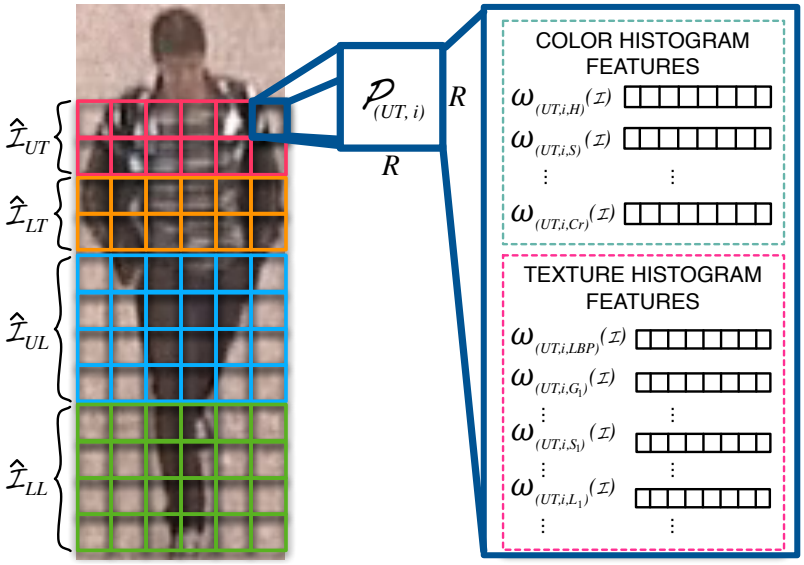


Figure 2.5: Dense image features from the detected body parts. Dense color and texture histogram features are extracted from each of the 4 resized body parts.

formation of any feature which, in turn, can be used for re-identification. In this work we focus not only on color features but also on popular texture features.

Before computing these features, we identify the salient regions like head $\mathcal{I}_H$, torso $\mathcal{I}_T$ and legs $\mathcal{I}_L$ from the given image $\mathcal{I}$ as proposed in [7]. In our approach we only consider $\mathcal{I}_T$ and $\mathcal{I}_L$ since the head region $\mathcal{I}_H$ often consists of a few and less informative pixels. We additionally divide both $\mathcal{I}_T$ and $\mathcal{I}_L$ into two horizontal sub-regions based on the intuition that people can wear shorts or long pants and short or long sleeves tops. The four different regions are resized to fixed height and width to extract fixed size dense features from all of them. We denote these resized regions as $\hat{\mathcal{I}}_\phi$ where $\phi \in \{UT, LT, UL, LL\}$ denotes the upper-torso, lower-torso, upper-legs and lower-legs regions respectively. The resized regions are further divided into non overlapping patches $\mathcal{P}_{(\phi,1)}, \mathcal{P}_{(\phi,2)}, \cdots, \mathcal{P}_{(\phi,n_\phi)}$ of size $R \times R$ each, where n_ϕ denotes the number of patches corresponding to the body part ϕ. Then, for all the patches $\mathcal{P}_{(\phi,i)}$, $i = 1, \cdots, n_\phi$ we extract the following features.

Color: State-of-the-art person re-identification methods use color features relying on the assumption that persons do not change their clothes as they move between camera FoVs.

According to that, and following the considerations in [72], we exploit the HSV, CIELab, RGB and YCbCr color spaces to extract the dense histogram features. For image $\mathcal{I}$, body-part ϕ and patch i we extract the histogram $\omega_{(\phi,i,c)}(\mathcal{I}) \in \mathbb{R}^{b_c}$, where b_c is the number of bins of the feature histogram for color component $c \in \{H, S, V, L, a^*, b^*, R, G, B, Y, Cb, Cr\}$.

Texture: Similar to color features, we extract dense texture features to capture the appearance of a person. We use LBP texture feature which is computationally efficient and is robust to both gray-scale variations [43] and rotation [85]. The extracted LBP texture histogram is denoted as $\omega_{(\phi,i,LBP)}(\mathcal{I}) \in \mathbb{R}^{b_{LBP}}$, where b_{LBP} is the number of bins used to quantize the resulting LBP histogram. We also use Gabor [36], Schmid [100] and Leung-Malik (LM) [68] filter banks to extract texture features. After convolving the i-th patch with each filter of the filter banks we compute the modulus of the response and quantize it in histograms of b_G, b_{Schmid} and b_{LM} bins respectively for the above 3 filter banks. Denoting the set of individual filters in Gabor, Schmid and LM filter banks as G, S and LM, the set of color and texture features extracted from patch $\mathcal{P}_{(\phi,i)}$ is given by the set $\{\omega_{(\phi,i,j)}(\mathcal{I})\}$ where $j \in \{c \cup LBP \cup G \cup S \cup LM\}$. An example of the responses of such filter banks is shown in the supplementary. Fig. 2.5 shows an example image where dense features from the 4 bodyparts have been extracted as described above.

2.4.2 Warp function space

To capture the transformation of the extracted features between cameras, we use the principles of Dynamic Time Warping (DTW). DTW [97] has been widely used in many fields such as speech recognition [51], data mining [53], activity recognition [106, 107] *etc.* DTW finds patterns that govern change of shape from one time series to another. This dynamic programming based algorithm non-linearly warps the time axis of a time series so that it is optimally aligned to the other time series with minimum cost of alignment. The

cost is, in general, the sum of the point to point-to-point distances of the two time series elements. Time sequences are functions of time while feature histograms are functions of the bin numbers. In our approach the bin number axis is warped to reduce the mismatch between feature values of two feature histograms from two cameras.

Let $\mathbf{x}(1, \cdots, m) = \langle x(1), \cdots, x(m) \rangle$ and $\mathbf{y}(1, \cdots, m) = \langle y(1), \cdots, y(m) \rangle$ be two vector valued functions. Let f be a warp function from $\mathbf{x}$ to $\mathbf{y}$, that is

$$y(a) = x(f(a)), f(a) : [1, m] \rightarrow [1, m] \in \mathcal{F} \tag{2.1}$$

where $\mathcal{F}$ is the space of all warp functions, the WFS.

To find the warp function, a cost matrix $C \in \mathbb{R}^{m \times m}$ is generated where the $(a, b)^{th}$ element (denoted as C_{ab}) of the matrix is given by the distance $\delta(x(a), y(b)), \forall a, b \in \{1, 2, \cdots, m\}$. Though any suitable distance function can be used or learned using a metric learning procedure, in general, the magnitude of the difference and the Euclidean distance between elements are adopted due to their simplicity [11]. The warp function is the path giving the lowest cumulative cost between fixed start point, the $(1, 1)^{th}$ cell and fixed end point, the $(m, m)^{th}$ cell of C. Let $\mathbb{W} = \{W_1, W_2, \cdots\}$ be the set of all possible paths between these two fixed points where W_q denotes the q^{th} path. W_q consists of tuples indicating the indices of the cells in C. Then the optimal warp path is given by,

$$W^* = \underset{W_q \in \mathbb{W}}{\operatorname{argmin}} \left(\sum_{(a,b) \in W_q} C_{ab} \right) \tag{2.2}$$

The optimization problem in (2.2) is solved in a dynamic programming framework under suitable monotonicity and continuity constraints [11, 83]. Finding the non-linear warp path W^* does not guarantee that the length of the warp path is same for all feature pairs $\mathbf{x}$ and $\mathbf{y}$. This is due to the fact that the mapping $f(a) : \{1, 2, \cdots m\} \rightarrow \{1, 2, \cdots m\}$, described by the tuples in W^* is, in general, many to many. To get a m length warp function

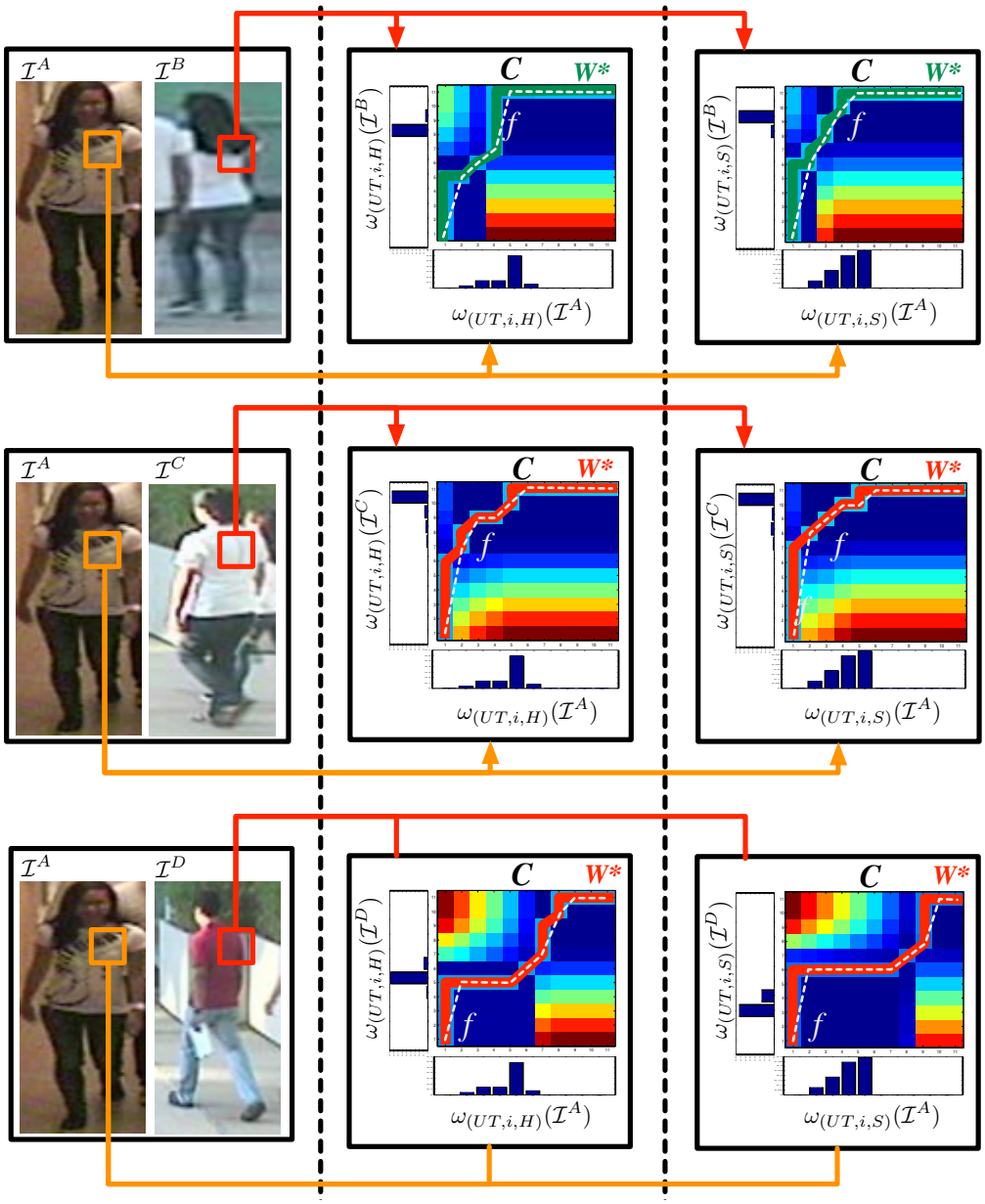


Figure 2.6: Example of computing the warp functions between features extracted from the same patch of two images. The first column shows two images from two cameras. The warp function between the features extracted from the same patches (shown by the orange and red boxes) are computed next. The last two columns show the cost matrices, the optimal warp path W^* and the corresponding warp function f. For convenience of visualization, warp functions computed for the H and S colorspaces only are shown in second and third column respectively. The cost matrix is colorcoded and the cost gets higher as the color goes from blue to red. First row shows the feature warps for the same person. Second and third rows show the warping of features between different persons that have similar and different appearance respectively with the person in the left.

we employ the following rule for all $(a, b) \in W^*$

$$f(a) = \begin{cases} \min(b) & \text{if } a \neq 1, m \\ a & \text{otherwise} \end{cases} \tag{2.3}$$

Gathering the $f(a)$'s for all $a = 1, 2, \cdots, m$ in a vector $\mathbf{f}_{(\mathbf{x},\mathbf{y})}(1, \cdots, m) = \langle f(1), \cdots, f(m) \rangle$ we get the warp function that warps $\mathbf{x}$ to $\mathbf{y}$.

In our approach the warp function $\mathbf{f}$ is computed for each feature and for every dense patch (see Section 2.4.1). In other words, as shown in Fig. 2.6, $\mathbf{f}$ is computed for feature pairs ($\omega_{(\phi,i,j)}(\mathcal{I}^A)$ and $\omega_{(\phi,i,j)}(\mathcal{I}^B)$) for each body part ϕ, patch i and feature j. The vector created by concatenating all such vector warp functions computed for the body part ϕ, is denoted as

$$\mathbf{F}_\phi(\mathcal{I}^A, \mathcal{I}^B) = \left\langle \mathbf{f}_{(\omega_{(\phi,i,j)}(\mathcal{I}^A), \omega_{(\phi,i,j)}(\mathcal{I}^B))} \right\rangle, \qquad \forall i, j \tag{2.4}$$

The set of all $\mathbf{F}_\phi(\mathcal{I}^A, \mathcal{I}^B)$'s computed between two images $\mathcal{I}^A$ and $\mathcal{I}^B$ of the same person forms the feasible or positive set $\mathcal{F}^p_\phi$ (for bodypart ϕ). The same computed between images of two different persons forms the infeasible or negative set $\mathcal{F}^n_\phi$. Both $\mathcal{F}^p_\phi$ and $\mathcal{F}^n_\phi$ together form the WFS which provides the description of the nonlinear feature transformations under different variabilities.

The proposed WFS model allows us to pose the re-identification problem as finding the parameters of the decision surface, that best separates the sets $\mathcal{F}^p_\phi$ and $\mathcal{F}^n_\phi$. Given a pair of candidate images, we classify such images as coming from the same target or not according as the warp functions between the image features lie in the positive or the negative region.

2.4.3 Re-identification in WFS

To re-identify persons moving across camera views we propose to train a binary classifier and classify the warp functions in the WFS as belonging to the feasible or infeasible sets. As discussed in Section 2.4.2 we use high-dimensional dense color and texture features to represent the appearance of the targets. While it is advantageous for a richer representation, it comes with the curse of dimensionality. The high dimensionality of the features result in high dimensional warp functions. Accordingly, any nonlinear classifier has to pay high computational and memory complexity in the training phase. This scalability issue makes it nontrivial to train a classifier directly on such high dimensional warp functions for large datasets whose training size is typically far beyond thousands. Therefore, we need to select a low dimensional subspace that can adequately handle the intrinsic dimensionality of the warp functions. Towards this objective, and supported by the recent study on real data discussed in [60], we use PCA [46] to embed the WFS into a low dimensional subspace. In the following we refer to $\mathbf{F}'_{\phi}(\mathcal{I}^A, \mathcal{I}^B)$ as the low dimensional warp function computed between images $\mathcal{I}^A$ and $\mathcal{I}^B$ for body part ϕ.

Even though PCA is able to reduce the dimensionality of the WFS, each dimension of it may not, still, be discriminating enough between the feasible and infeasible warp functions. Thus a classifier giving more importance to the more discriminative dimension is desirable. A random forest (RF) [14] is a popular and efficient classifier based on bootstrapped aggregation ideas. It is a combination of many binary decision trees built using several bootstrap samples. At each node of each tree a subset of the warp function dimensions is randomly chosen and the best split is calculated only within this subset. This randomization of the warp function dimensions effectively chooses the dimensions according to their importance in separating the feasible and the infeasible warp functions in the WFS.

This coupled with the reduction of overfitting error makes RF a suitable choice to learn the parameters of the decision boundary.

In the classification phase the warp function between the features of two candidate images from two different cameras is computed. The trained RF classifies the warp function as coming from the same target or not according as it lies in the positive or the negative region.

Let $\mathcal{I}^{A_1}, \cdots, \mathcal{I}^{A_N}$ be the N images of a given person $\mathcal{A}$ and $\mathcal{I}^{B_1}, \cdots, \mathcal{I}^{B_M}$ be the M images of another person $\mathcal{B}$ in another camera. As commonly accepted in the field of person re-identification, if N=1 and M=1, then the approach is defined to be a *single-shot* approach, otherwise, if both N and M are greater than 1, it is named a *multiple-shot* approach. As the total number of possible warp functions that can be computed for a single body part ϕ is $N \times M$, we have $|\phi| \times N \times M$ predicted probabilities for a target pair, where $|\phi|$ denotes the number of parts into which the body of a person is divided. The probability of $\mathcal{A}$ and $\mathcal{B}$ being the same person is computed by averaging all the $|\phi| \times N \times M$ probabilities obtained from the classifier.

2.5 Experiments

We evaluated our approach on five publicly available datasets, the ETHZ dataset [31], the CAVIAR4REID dataset [17], the VIPeR dataset [41], the WARD dataset [80] and a dataset (RAiD) [21] collected by us. We chose these datasets because they provide many challenges faced in real world person re-identification applications, *e.g.*, viewpoint, pose and illumination changes, different backgrounds, image resolutions, occlusions, *etc.* Of these, WARD and RAiD are specifically geared towards large illumination change. More details about each dataset are reported in Table 2.1 and are discussed below. We report the re-

Table 2.1: Details and comparison of commonly used person re-idenentification benchmark datasets. For the CAVIAR4REID dataset, values in brackets are for persons appearing in both cameras. For ETHZ dataset values in brackets are for SEQ.#1, SEQ.#2 and SEQ.#3 respectively.

Dataset	People	Image info	Cameras	Additional Info
ETHZ [101] (SEQ.#1,SEQ.#2, SEQ.#3)	(83, 35, 28)	Images: (4856, 1690, 1762) Avg. images per person per camera: (59, 48, 63) Size: 13×30 to 158×432	(1,1,1)	Scenario: outdoor Challenges: color changes, occlusions, sptial resolution http://homepages.dcc.ufmg.br/~william/
CAVIAR4REID [17]	72 (50)	1220 (1000) Avg. images per person per camera: 10 (10) Size: 17×39 to 72×144	2	Scenario: indoor Challenges: viewpoint variation, color changes, spatial resolution www.lorisbazzani.info
WARD [80]	70	Images: 4786 Avg. images per person per camera: 69 Size: 15×36 to 70×189	3	Scenario: outdoor Challenges: viewpoint variations, spatial resolution, color changes http://users.dimi.uniud.it/~niki.martinel/
VIPeR [41]	632	Images: 1264 Avg. images per person per camera: 1 Size: 48×128	2	Scenario: outdoor Challenges: viewpoint variation, color changes http://vision.soe.ucsc.edu/node/178
RAiD [21]	43	Images: 6920 Avg. images per person per camera: 40 Size: 64×128	3	Scenario: outdoor and indoor Challenges: Severe illumination and viewpoint variations, spatial resolution changes http://www.ee.ucr.edu/~amitrc/datasets.php

sults for both single-shot (N=1) and multiple-shot (N >1) strategies. For all multiple-shot strategies we use N=M. Results are shown in terms of recognition rate as Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values, as commonly performed in the literature. The CMC curve is a plot of the recognition percentage versus the ranking score and represents the expectation of finding the correct match inside top k matches. On the other hand, nAUC gives an overall score of how well a re-identification method performs irrespective of the dataset size. For each dataset the evaluation procedure is repeated 10 times using independent random splits. We reported the average results on these 10 splits. All the results used for comparison were either taken from the corresponding works or by running the publicly available codes on datasets for which reported results could not be obtained. We did not re-implement other methods as it is very difficult to exactly emulate all the implementation details.

2.5.1 Implementation Details

In our implementation we used the following settings:

- Image pairs of the same or different person(s) in different cameras were randomly picked to compute the positive and negative warp functions respectively;
- $\hat{\mathcal{I}}_{UT}$, $\hat{\mathcal{I}}_{LT}$, $\hat{\mathcal{I}}_{UL}$ and $\hat{\mathcal{I}}_{LL}$ have been resized as follows:
  - For the ETHZ dataset: $\hat{\mathcal{I}}_{UT} = \hat{\mathcal{I}}_{LT} = \hat{\mathcal{I}}_{UL} = \hat{\mathcal{I}}_{LL} = 32 \times 16$;
  - For the CAVIAR, WARD and RAiD dataset: $\hat{\mathcal{I}}_{UT} = \hat{\mathcal{I}}_{LT} = \hat{\mathcal{I}}_{UL} = \hat{\mathcal{I}}_{LL} = 64 \times 32$
  - For the VIPeR dataset: $\hat{\mathcal{I}}_{UT} = \hat{\mathcal{I}}_{LT} = \hat{\mathcal{I}}_{UL} = \hat{\mathcal{I}}_{LL} = 48 \times 32$;
- The size of each dense patch has been selected to be $R \times R = 8 \times 8$ pixels.
- The color histograms extracted from the dense patches were quantized using $b_c = 10$ bins for each color space component c.

- Texture features have been extracted using the following parameters:
  - LBP: we followed the same protocols used in [85]. LBP histograms were quantized into $b_{LBP} = 10$ bins.
  - Gabor: we used Gabor filters at 8 orientations and 5 scales. b_G was set to 16.
  - Schmid: the same filter settings as [100] have been used. b_{Schimd} was set to 16.
  - Leung-Malik: the same filter bank defined in [68] consisting of 36 oriented filters with 6 orientations, 3 scales and 2 phases, 8 Laplacian of Gaussian (LoG) filters, and 4 Gaussians was used. b_{LM} was set to 16.
- δ was taken as the Euclidean distance between the feature values.
- While doing PCA, we selected the largest principal components such that the 99% of the original variance is retained.
- The RF parameters such as the number of trees, the number of features to consider when looking for the best split, *etc* were selected using 4-fold cross validation.

2.5.2 Comparative Evaluation on Benchmark Datasets

The proposed method is, first, evaluated on 3 challenging benchmark datasets, namely ETHZ, CAVIAR4REID and VIPeR. Since WARD and RAiD contain a large illumination variation, we show the performance on these two datasets separately in the next sub-section.

ETHZ Dataset

The ETHZ dataset [31] contains video sequences of urban scenes captured from moving cameras. It contains a large number of different people in uncontrolled conditions. It has originally been proposed for pedestrian detection, but in [101] a modified version of the

dataset was provided for the task of person re-identification. This version consists of person images extracted from three video sequences structured as follows: SEQ. #1 containing 83 persons (4,857 images), SEQ. #2 containing 35 persons (1,961 images), and SEQ. #3 containing 28 persons (1,762 images). Since the original video sequences are captured from moving cameras, images have a range of variations in human appearance and some even suffer from heavy occlusions. However, for the same reason the dataset does not provide a realistic scenario for person re-identification with multiple disjoint cameras. To make this dataset more challenging, we followed the strategy proposed in [6] by randomly picking a set of 10 consecutive frames from the beginning and from the end of each sequence.

Despite this limitation it is commonly used for person re-identification, so we also evaluated our approach on this dataset. Following the evaluation setup in [101, 7], all images have been resized to 32×64 pixels. We evaluate our method using both single-shot and multiple-shot strategies. Similar to [44, 45], for the single-shot scenario, we randomly sample two images per person to build a training set, and another two images to build the test set. The test images from one camera constitute the probe and the those from the other camera create the gallery set.

In Table 2.2 we present the performance of our method using both single-shot and multiple-shot strategies. The first 9 rows show the performance comparison with 8 different methods when 1 single image has been used to build the gallery and the probe sets. The last 10 rows show the performance comparison with 9 different methods using a multiple-shot strategy. For the single shot scenario our performance is either superior to or same with that of all the 8 methods for each of the 3 sequences. For the multiple-shot scenario the same settings of experiments as in [117, 77] were used with N=5. In this scenario, the BRM [6] approach has superior performances only from rank 1 to rank 4 for SEQ.#1 . Similarly the eLDFV [78] method has superior performance compared to our method for rank 1 to 3.

Our method is the only one that achieves the 99% of correct recognition for this sequence within the top 7 rank scores. On SEQ.#2 we outperform all other methods as we reach 100% correct recognition within top 4 matches. Similarly, on SEQ.#3 our method has the best performance and recognizes all the persons at rank 1. Notice that in these experiments we are using N=5 images, whereas the results for SDALF, AHPE, eBiCov and BRM were reported using N=10 images. For all the three sequences in the ETHZ dataset our method is the only one that achieves the 99% of correct recognition within the top 7 matches.

CAVIAR4REID Dataset[17]

This dataset [17] contains images of pedestrians extracted from the CAVIAR repository. It is composed of 1220 images of 72 pedestrians out of which 50 are viewed by two disjoint cameras. So, in our approach we considered only these 50 persons. It is more interesting than the ETHZ, where images are extracted from a single camera. Other challenges in this dataset includes a broad change in the image resolution, with a minimum and maximum size of 17×39 and 72×144, respectively, severe pose variations, illumination changes and occlusion.

It is common to split the CAVIAR4REID dataset both in terms of people [5, 91] and not [7, 69]. We conducted experiments following both these protocols to fairly compare against methods following either of these two. Following the same setup as in [5] first, the 50 people are equally divided into training and test sets of 25 persons each. In this setup we compare against LF [91] and ICT [5] who use a multiple shot strategy with N=5 and N=10 images respectively. In Fig. 2.7(a) we show that our algorithm outperforms both the methods and reaches as high as 40.9% rank 1 score when a multiple shot strategy with N=10 is employed. In the second set up following the same protocol as in [69], we do not split

Table 2.2: Comparison of the proposed method on the ETHZ dataset using both a single shot-strategy (top 9 rows) and a multiple-shot strategy (last 10 rows). Recognition rates for top 7 ranks are shown for each of the three sequences. The best recognition rates for each rank are shown in boldface font

Method	SEQ.#1							SEQ.#2							SEQ.#3						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Proposed (1 image)	**84**	**88**	**91**	**93**	**94**	**95**	**96**	**81**	**86**	**90**	**93**	**95**	**96**	**97**	**91**	**97**	**99**	**99**	**99**	**99**	**100**
eLDFV [78](1 image)	83	87	90	91	92	93	94	79	84	87	90	91	92	93	**91**	94	96	97	97	97	97
SDALF [7](1 image)	65	73	77	79	81	82	84	64	74	79	83	85	87	89	76	83	86	88	90	92	93
eBiCOV [77](1 image)	74	80	83	85	87	88	89	71	79	83	86	88	90	91	82	87	90	92	93	94	95
eSDC_knn [118](1 image)	81	86	89	90	92	93	94	79	84	87	90	91	92	93	90	95	96	97	98	98	99
eSDC_ocsvm [118](1 image)	80	85	88	90	91	92	93	80	**86**	89	91	93	94	95	89	94	96	97	98	98	99
RPLM [44](1 image)	77	83	87	90	91	92	92	65	77	81	82	86	89	90	83	90	92	94	96	96	97
IBML [45](1 image)	78	84	87	89	90	91	91	74	81	84	87	89	91	92	**91**	95	97	98	98	98	99
ICT [5](1 image)	68	76	82	86	87	89	90	70	82	89	91	93	94	95	**91**	94	96	97	97	98	98
Proposed (5 images)	94	95	96	97	**98**	**98**	**99**	**98**	**99**	**99**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
PLS [101](all images)	79	85	86	87	88	89	90	74	79	81	83	84	85	87	77	81	82	84	85	87	89
eBiCOV [77](5 images)	93	94	95	95	96	96	96	91	94	95	96	97	97	97	98	98	99	**100**	**100**	**100**	**100**
eLDFV [78](5 images)	**96**	**97**	**97**	97	**98**	**98**	98	97	98	**99**	**100**	**100**	**100**	**100**	99	**100**	**100**	**100**	**100**	**100**	**100**
LDC [117](5 images)	92	95	96	97	**98**	**98**	98	92	95	97	98	99	99	99	96	97	98	99	99	99	99
ICT [5](5 images)	92	93	94	95	96	96	97	95	98	**99**	99	**100**	**100**	**100**	95	96	97	99	**100**	**100**	**100**
SDALF [7](10 images)	91	92	93	94	94	94	94	91	94	96	96	97	97	98	94	95	96	96	96	96	96
AHPE [8] (10 images)	85	89	92	93	94	94	95	80	86	89	92	93	94	95	83	91	92	94	96	97	97
eBiCOV [77](10 images)	93	94	95	96	96	96	96	91	95	96	97	98	99	99	97	98	99	**100**	**100**	**100**	**100**
BRM [6](10 images)	**96**	**97**	**98**	**98**	**98**	**98**	98	94	95	95	95	95	95	96	96	98	**100**	**100**	**100**	**100**	**100**

the dataset in terms of persons. Pairs of images are randomly selected in different views for training. The probe and the gallery sets are formed by randomly selecting images from the remaining ones for each person. In this scenario we compare against the methods who have adopted the same strategy of split. Namely the methods are AHPE [8], SDALF [7], CI [59], CPS [17], LAFT [69] and LDC [117]. Fig. 2.7(b) shows the CMC curves for the single shot scenario. Fig. 2.7(c) and (d) show the comparison with the multi-shot strategy. While for single shot scenario we meet the state-of-the-art performance of LAFT and outperform the rest, for both the multishot scenarios we have superior performance over all the compared methods.

VIPeR Dataset

VIPeR [41] is a challenging dataset for person re-identification due to the changes in illumination and pose, and the low spatial resolution of images. This dataset contains one image each from two cameras of 632 persons. Although images from the same camera are not always taken from the same viewpoint and thus do not fully fit our framework, still we compare our results with other methods to show that the proposed approach achieves good results in such a scenario too. To evaluate our method we followed the same normalization approach as in [7, 5, 118], resizing all the images to 48×128 pixels. To compare our approach to state-of-the-art methods we used the same evaluation protocol proposed in [42]. We split the dataset in terms of persons and used 316 of them for training and the remaining 316 for testing. As the VIPeR dataset is a single-shot dataset, we used N=1 images per person to form the training and test sets.

In Table 2.3 we report the recognition performance for the top 100 ranks and compared the results with 20 state-of-the-art methods for person re-identification. The

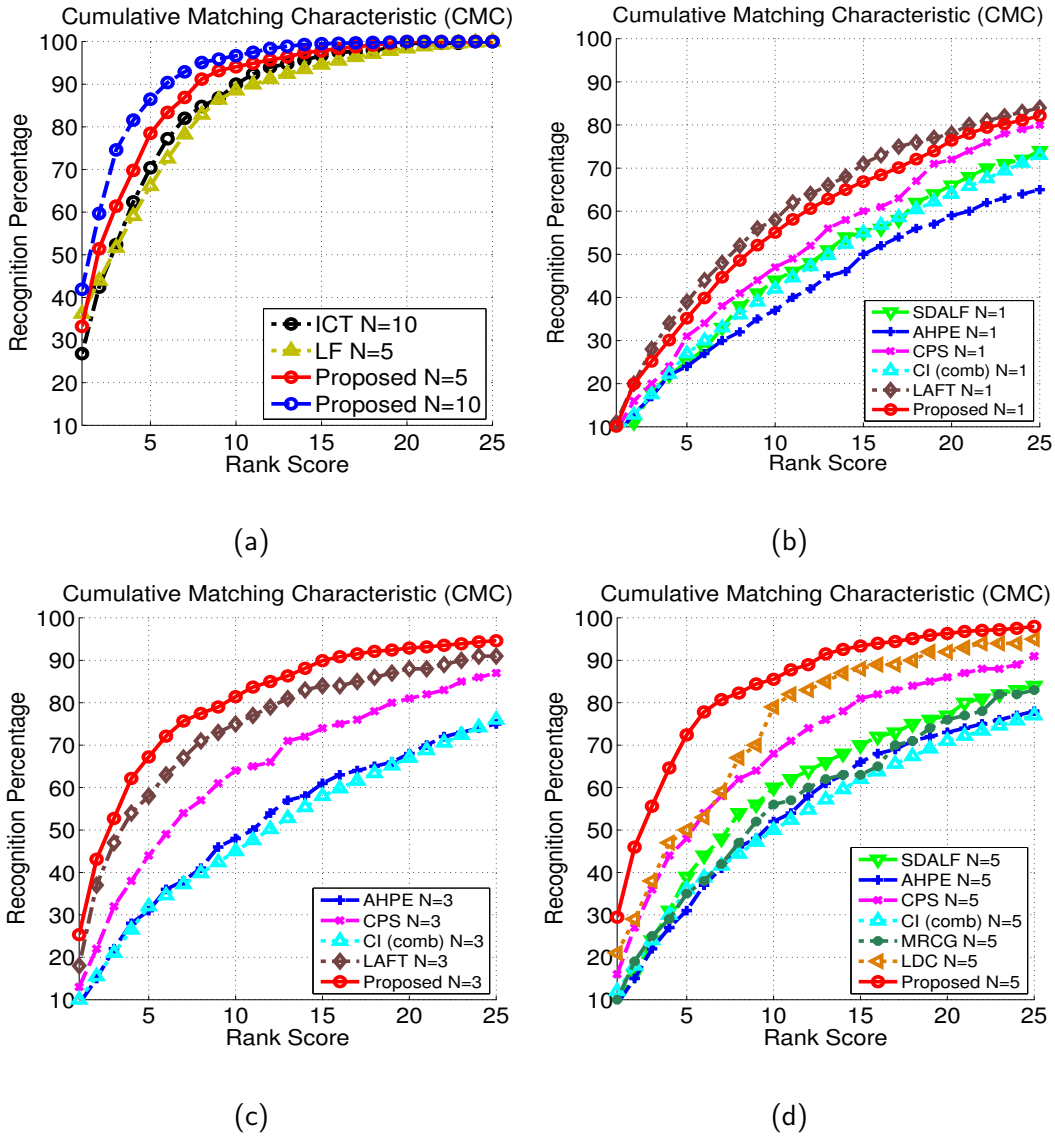


Figure 2.7: CMC curves for CAVIAR4REID dataset. In (a) results are shown when the dataset is split in terms of persons. In (b), (c) and (d) comparisons are shown for the case where the dataset is not split in terms of persons with N=1, N=3 and N=5 respectively.

Table 2.3: Comparison of the proposed method on the VIPeR dataset. Top 100 rank matching rate (percent) is shown.

Rank Score	1	10	20	50	100
Proposed	25.81	69.56	83.67	95.12	98.89
RCCA [3]	**30.00**	**75.00**	**87.00**	96.00	**99.00**
LAFT [69]	29.60	69.30	81.34	**96.80**	**99.00**
LF [91]	24.18	67.12	81.38	94.12	
TML [70]	19.00	61.00	74.00	91.00	97.00
KISSME [57]	19.60	62.20	74.92	91.80	98.00
RPLM [44]	27.00	69.00	83.00	95.00	**99.00**
IBML [45]	22.00	63.00	78.00	93.00	98.00
ELF [42]	12.00	43.00	60.00	81.00	93.00
SDALF [7]	19.87	49.73	65.73	84.80	
PRSVM [95]	14.60	53.90	70.10	85.00	94.00
CPS [17]	21.84	57.21	71.00	88.10	
PRDC [120]	15.70	53.86	70.09	87.00	
LMNN-R [26]	23.70	68.00	80.00	93.00	**99.00**
eBiCOV [77]	20.66	56.18	68.00	84.90	
eLDFV [78]	22.34	60.04	71.00	88.92	**99.00**
eSDC.knn [118]	26.31	58.86	72.77	79.30	
eSDC.ocsvm [118]	26.74	62.37	76.36	82.10	
CI [59]	18.00	50.00	62.00	81.00	
ICT [5]	15.90	57.20	78.30	91.00	95.00
ARLTM [73]	21.00	52.00	68.00	86.00	

table shows that the proposed method does achieve a performance better than most of the state-of-the-arts as far as the performance corresponding to rank 1 is considered. It is behind the top performer only by 4.19% for rank 1. The performance continuously improves with higher ranks. The rank 100 performance is either same or better than all the methods. According to [5] the performance at higher ranks is, sometimes, more significant as this reflects the algorithm's performance for difficult cases. Thus, in this challenging dataset with only one image per person in two non-static cameras the proposed method does achieve competitive performance as that of the state-of-the-arts.

2.5.3 Comparative Evaluation with Large Appearance Variation

Since our focus is to understand the space of transformation of features, we provide the performance of the proposed method for 2 datasets which posses significant appearance variation.

WARD Dataset

The WARD dataset [80] contains 4786 images of 70 different people acquired by three non-overlapping cameras in a real surveillance scenario. This dataset is of particular interest because it has a huge illumination variation apart from resolution and pose changes. We conducted the experiments for all the three different camera pairs, denoted here as camera pairs 1-2, 1-3, and 2-3. The proposed approach is compared with the methods for which either the CMC performance on this dataset is reported in literature or the code is available. Namely the methods are SDALF [7], WACN [80] and ICT [5]. Fig. 2.8(a), (b) and (c) compare the performance adopting a multiple shot strategy with N=10 for camera pairs 1-2, 1-3, and 2-3, respectively. The 70 people in this dataset are equally divided into training and test sets of 35 persons each. For all 3 camera pairs the proposed method outperforms the rest with rank 1 recognition percentage as high as 51.6% for the camera pair 2-3. The next runner up has the recognition percentage of 29.5% for rank 1. For all the camera pairs 97% recognition performance is reached within top 10 matches.

RAiD Dataset

This dataset was collected with a view to have large illumination variation that is not present in most of the publicly available benchmark datasets. In the original dataset 43 subjects were asked to walk through 4 cameras of which 2 are outdoor and 2 are indoor

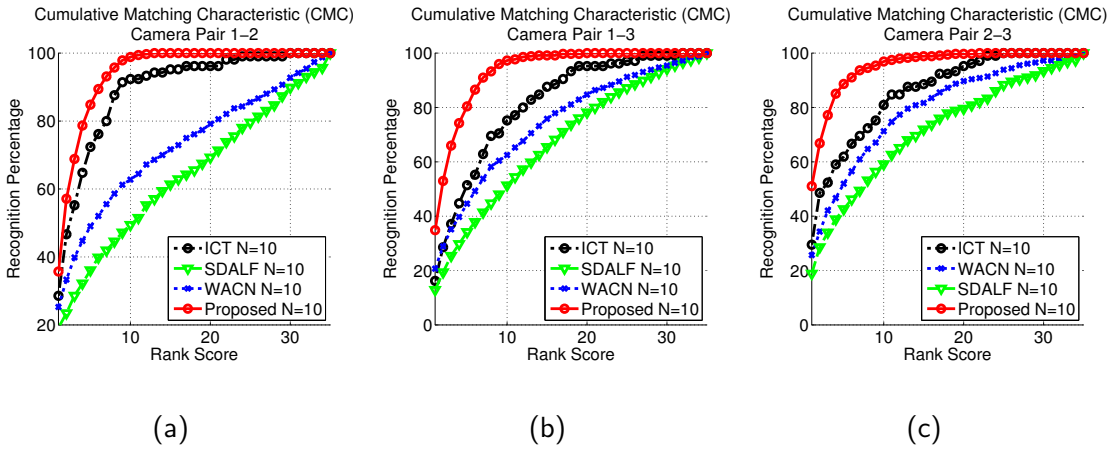


Figure 2.8: CMC curves for the WARD dataset. Results and comparisons in (a), (b) and (c) are shown for the camera pairs 1-2, 1-3, and 2-3 respectively. All the results are reported for the case where the dataset is split in terms of persons with N=10.

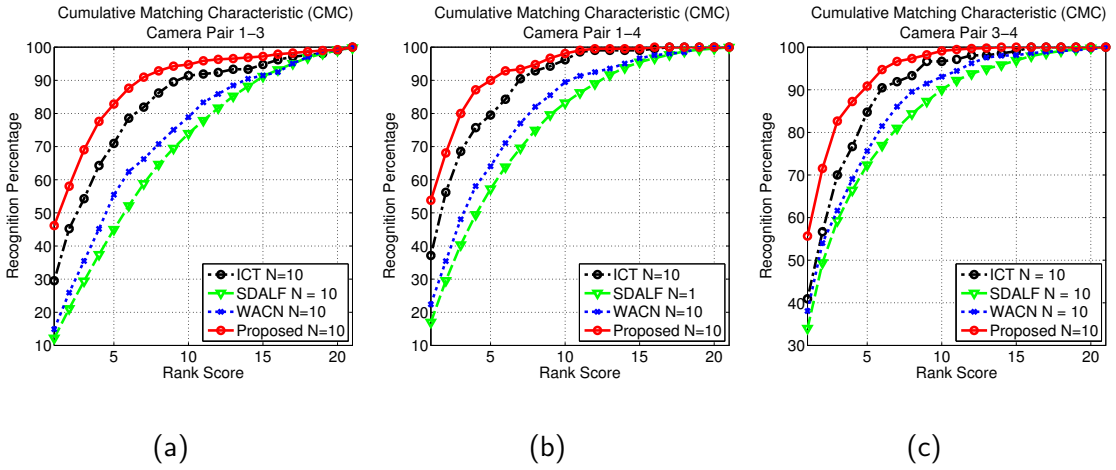
Figure 2.9: Sample images of persons from the RAiD dataset showing the variation of appearance between the indoor and the outdoor cameras.

to make sure there is enough variation of appearance between cameras. To reduce the number of pairs of cameras and yet to keep the variation of light to maximum we chose to experiment with 3 of these cameras, 1 indoor and 2 outdoors. These 3 cameras contain 6060 images of 41 persons walking through 1 indoor (denoted as camera 1) and 2 outdoor cameras (denoted as camera 3 and camera 4). Sample images showing the variation of illumination between the cameras are shown in Fig. 2.9.

The proposed approach is compared with the methods for which the code is available. Namely the methods are WACN [80], SDALF [7] and ICT [5]. The dataset was split in terms of persons with 22 persons forming the training set and the rest 21 persons forming the test set. Fig. 2.10(a), (b) and (c) compare the performance adopting a multiple shot strategy with N=10 for camera pairs 1-3, 1-4 and 3-4 respectively. We see that the proposed method is superior to all the rest for both the cases when there is not much appearance variation (camera pair 3-4) and when there is significant lighting variation (for camera pairs 1-3 and 1-4). Expectedly, for camera pair 3-4 the performance is the best achieving 55.7% rank 1 performance. For the other two difficult cases too, the proposed method is superior to all the rest achieving 46.4% and 53.9% rank 1 performances for camera pairs 1-3 and 1-4 respectively. The second best performance is that of ICT which achieves 29.5% and 37.3% rank 1 performances for camera pairs 1-3 and 1-4 respectively. Fig. 2.11 shows a comparison of re-identification performances with ICT [5] (achieving the next best performance). The comparison is done on 10 randomly selected persons. For viewing convenience only the top 15 candidates are shown. The green bounding box highlights the ground truth match for each of the query persons. The ground truth match is within top 3 ranked matches for 9 out of the 10 examples while 6 out of these 10 persons are the highest ranked matches too. For the same set of persons the ground truth match is within top 3 ranked matches for 2 out of the 10 examples in ICT. None of them is the highest ranked match.

2.5.4 Average Performance across Multiple Datasets

Having shown the performance of the proposed method on separate datasets with different challenges, in this sub-section we show that the proposed method gives the most consistent performance across different datasets each having multiple different challenges.



(a) (b) (c)

Figure 2.10: CMC curves for RAiD dataset. In (a), (b) and (c) comparisons are shown for the camera pairs 1-3, 1-4 and 3-4 respectively.

The performance is measured in terms of average nAUC values across different combinations of the 4 publicly available benchmark datasets (ETHZ, WARD, CAVIAR4REID and VIPeR). We compare with 14 state-of-the-art methods for which either the code is available or results for at least 2 of these 4 datasets are reported. The nAUC values for different methods are either taken from the reported results or computed from the reported CMC curves. To make a fair comparison we consider all combinations of 2 or more datasets and compare our performance by averaging over the datasets separately for each combination. Table 2.4 shows the performance comparison. The proposed method has the highest average nAUC value for 10 out of the 11 possible combinations. The only case (combination of ETHZ and CAVIAR) where the proposed method is the runner up, the nAUC value changes only at the 3^{rd} decimal place. The superior performance of the proposed method on any combination of these datasets establishes the fact that the proposed method is not tuned to any specific dataset and can address varied number of challenges across different datasets better than the state-of-the-art.

Table 2.4: Comparison of average performance across different datasets

# of datasets	4	3				2					
	ETHZ WARD CAVIAR VIPeR	ETHZ WARD CAVIAR	ETHZ WARD VIPeR	ETHZ CAVIAR VIPeR	WARD CAVIAR VIPeR	ETHZ WARD	ETHZ CAVIAR	ETHZ VIPeR	WARD CAVIAR	WARD VIPeR	CAVIAR VIPeR
Proposed	**0.9377**	**0.9292**	**0.9666**	**0.9352**	**0.9200**	**0.9683**	0.9211	**0.9772**	**0.8983**	**0.9544**	**0.9072**
ICT [5]	0.9115	0.8977	0.9302	0.9298	0.8883	0.9188	0.9182	0.9669	0.8561	0.9048	0.9042
SDALF [7]	0.8230	0.7898	0.8538	0.8697	0.7786	0.8195	0.8433	0.9393	0.7066	0.8026	0.8264
RPLM [44]	-	-	-	-	-	-	-	0.9566	-	-	-
IBML [45]	-	-	-	-	-	-	-	0.9549	-	-	-
CPS [17]	-	-	-	0.9062	-	-	0.8914	0.9674	-	-	0.8600
eBiCOV [77]	-	-	-	-	-	-	-	0.9394	-	-	-
eLDFV [78]	-	-	-	-	-	-	-	0.9622	-	-	-
eSDC.knn [118]	-	-	-	-	-	-	-	0.9335	-	-	-
eSDC.ocsvm [118]	-	-	-	-	-	-	-	0.9402	-	-	-
LDC [117]	-	-	-	-	-	-	**0.9250**	-	-	-	-
AHPE [8]	-	-	-	-	-	-	0.8245	-	-	-	-
LAFT [69]	-	-	-	-	-	-	-	-	-	-	0.8820
LF [91]	-	-	-	-	-	-	-	-	-	-	0.8980
CI (comb) [59]	-	-	-	-	-	-	-	-	-	-	0.7948

Figure 2.11: Visual comparison of matches using feature warps for camera pair 1-3 of the RAiD dataset. First column is the probe image. Second and third columns show the top 15 matches computed using the proposed method and ICT [5] respectively.

2.5.5 Robustness to Choice of Classifiers and Patch Size Parameters

To further test the robustness of the proposed method to the choice of classifiers, experiments were conducted with another classifier, namely a Support Vector Machine (SVM) [16]. In a similar way, the proposed method is run with different values of another critical parameter, the patch size of the dense features. We ran these experiments with two datasets, namely WARD and RAiD. In Table 2.5 we report the recognition performance for different choices of these parameters in terms of the nAUC values. For different choices of the classifiers or for different patch sizes, all the other parameters are chosen as described in Section 2.5.1.

Table 2.5: Comparison of performance for different choices of classifiers and patch sizes

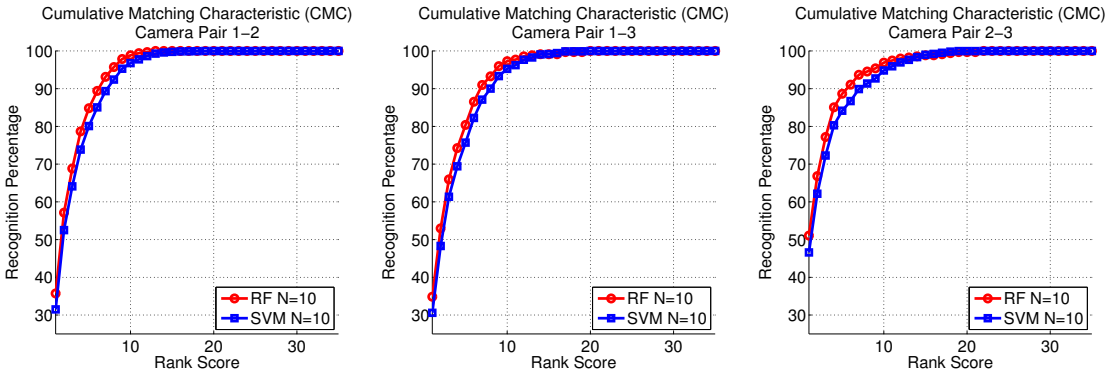
Dataset	Camera pair	Classifiers		Patch size		
		RF	SVM	4×4	8×8	16×16
WARD	1-2	0.9437	0.9313	0.8996	0.9437	0.9302
	1-3	0.9386	0.9268	0.8896	0.9386	0.9207
	2-3	0.9542	0.9426	0.9081	0.9542	0.9394
RAiD	1-3	0.8905	0.8755	0.8296	0.8905	0.8754
	1-4	0.9295	0.9122	0.8670	0.9295	0.9123
	3-4	0.9395	0.9216	0.8771	0.9395	0.9220

Performance comparison for different choices of classifiers

Here we provide the comparison of re-identification performance in terms of the CMC curves as different classifiers are used. Following the same convention as used throughout the chapter, the patch size used for both the classifiers, is 8×8. Fig. 2.12 and 2.13 show the CMC curves showing the comparison of re-identification performance with two different classifiers (RF and SVM) for WARD and RAiD dataset respectively. In Table 2.5 and in the plots provided in fig. 2.12 and 2.13, it is shown that the performance is similar even if the classifier is changed to an SVM for both the datasets. As shown in Table 2.5 the nAUC values differ only at the second decimal places for all the camera pairs with a maximum change of 0.0179 for camera pair 3-4 of the RAiD dataset.

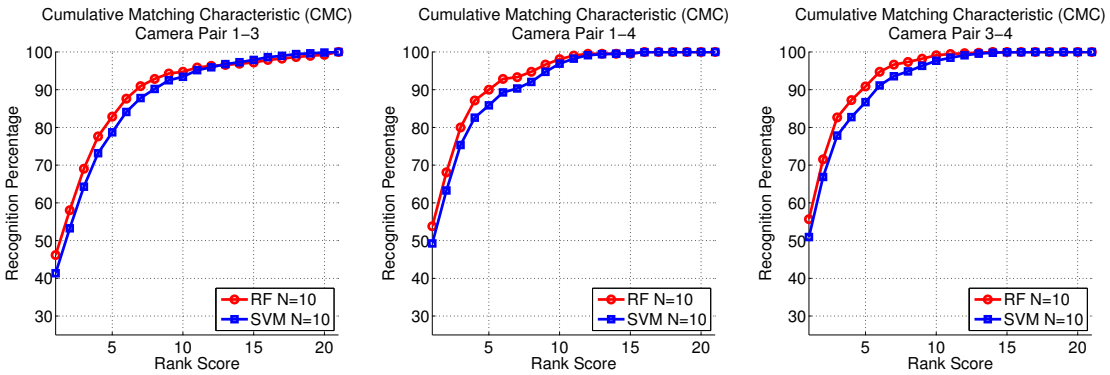
Performance comparison for different choices of patch sizes

Here we provide the comparison of re-identification performance in terms of the CMC curves as three different dense feature patch sizes ($4 \times 4, 8 \times 8$ and 16×16) are used. A RF classifier is chosen for the experiments with these three different patch sizes. All the other parameters are chosen as described in Section 2.5.1. Fig. 2.14 and 2.15 show the CMC curves showing the comparison of re-identification performance with these three



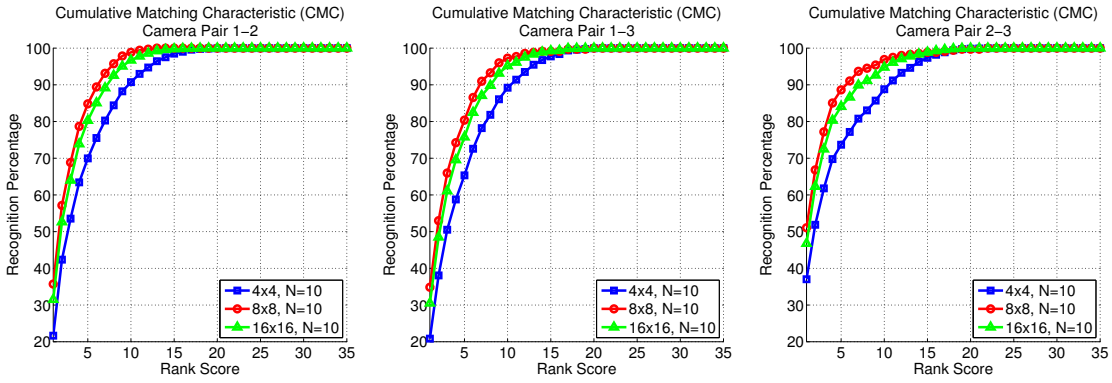
(a) (b) (c)

Figure 2.12: CMC curves showing the comparison of re-identification performance with two different classifiers (RF and SVM) for WARD dataset. In (a), (b) and (c) comparisons are shown for the camera pairs 1-2, 1-3 and 2-3 respectively.



(a) (b) (c)

Figure 2.13: CMC curves showing the comparison of re-identification performance with two different classifiers (RF and SVM) for RAiD dataset. In (a), (b) and (c) comparisons are shown for the camera pairs 1-3, 1-4 and 3-4 respectively.



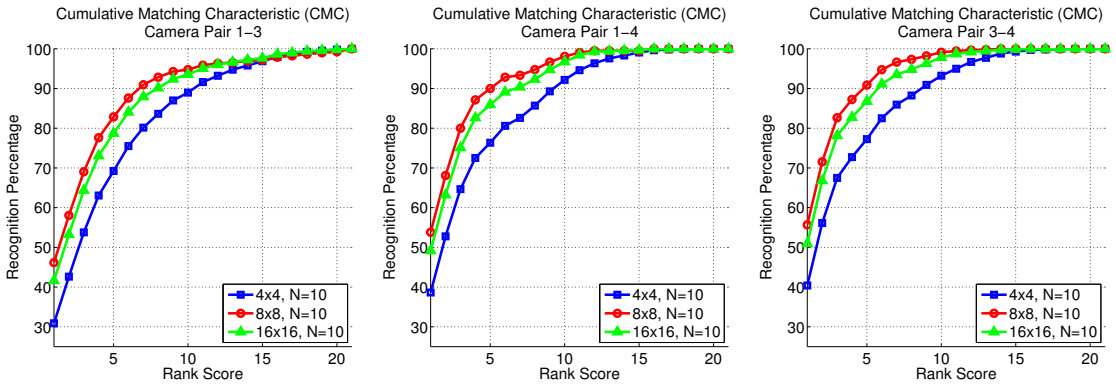
(a) (b) (c)

Figure 2.14: CMC curves showing the comparison of re-identification performance with three different dense patch sizes ($4 \times 4, 8 \times 8$ and 16×16) for WARD dataset. In (a), (b) and (c) comparisons are shown for the camera pairs 1-2, 1-3 and 2-3 respectively.

different dense feature patch sizes for WARD and RAiD dataset respectively. Similar to different choices of classifiers, no major change of the performance is noted for the 3 different settings of patch sizes for which we conducted the experiment. Indeed the change in the nAUC values is in the second decimal place also for different choice of dense patch sizes with the best performance being observed by a patch size of 8×8. This establishes the robustness of the proposed method to the choice of different classifier types and the dense feature patch sizes.

2.6 Conclusions

In this work we addressed the problem of multi-camera target re-identification by finding a nonlinear warp function between features from two cameras. Given a pair of feature vectors we show that we can learn the decision surface best separating the feasible and infeasible set of warp functions in the WFS. The target re-identification problem is posed as classifying a test warp function as belonging to the set of feasible or infeasible warp



(a) (b) (c)

Figure 2.15: CMC curves showing the comparison of re-identification performance with three different dense patch sizes ($4 \times 4, 8 \times 8$ and 16×16) for RAiD dataset. In (a), (b) and (c) comparisons are shown for the camera pairs 1-3, 1-4 and 3-4 respectively.

functions. We show that our approach is robust with respect to severe illumination and pose variations by evaluating the performance on five datasets. Our approach outperforms the existing state-of-the-art methods for person re-identification. The future directions of our research will be to apply our approach to capture the transformation of more complex features and to study its application for multi-target tracking in a non-overlapping multi-camera scenario.

Chapter 3

Person Re-identification through Sparse Non-redundant Representative Selection

In this chapter we address person re-identification from a continuous learning perspective. In particular we involve human in the loop to help boosting re-identification performance and at the same time keep the costly human labor to a minimum. Starting with a completely unlabeled pool of images, we choose a small representative subset of the whole unlabeled pool to be given to the human annotator for labeling so that maximum information is gained by annotating just a few difficult representative examples. In addition to choosing only a few exemplars to reduce human annotation effort the proposed approach exploits redundancy in the unlabeled pool so that the chosen exemplars are diverse in the sense that the human annotator does not have to label images of the same person repetitively. We adopt a convex optimization based strategy towards this objective in an iterative framework. We also use a structure preserving sparse reconstruction based

classifier to reduce the training burden typically seen in discriminative classifiers. The two stage framework not only helps in reducing the labeling effort but also can handle situations when new unlabeled data arrives continuously. This is due to the fact that online update of the classifier involves only the incorporation of new labeled data rather than any expensive training phase. Using three benchmark datasets, we validate our approach and demonstrate that our framework achieves superior performance with significantly less amount of manual labeling.

3.1 Introduction

Traditional person re-identification techniques rely on an intensive supervised training phase where several images of different persons are assumed to be labeled beforehand. Apart from high cost of labeling the training data, all the data may not be available at the very outset. Moreover, a static pre-trained model can not adapt to the changing dynamics of the incoming data. The traditional approaches try to capture large appearance variations of different persons across cameras by labeling as many images of them as possible. Considering the time, labor and human expertise involved in labeling the training data manually, person re-identification for a large number of persons, often, suffers from the curse of scalability when using traditional but otherwise tested approaches.

Active learning [102] is a natural choice for reducing labeling efforts by asking for labels only on a few but informative samples (called the active samples), rather than seeking labels either for samples chosen randomly from a set or for the whole set. With the unprecedented data deluge of the current age, a relevant question has been whether or not the quest for more data does contribute to improved performance [121, 63]. Keeping this question aside, we argue that, in order to truly reduce the labeling cost we need to choose

a sparse but informative set of samples to be labeled. As only a small part of the whole data is annotated, the annotation effort is reduced considerably compared to annotating the whole dataset.

In presence of multiple cameras with varying angle of views, scale, resolution and illumination condition, the appearance of a person captured in each camera can have significant variations. Active learning based methods [15, 30, 50], though, have been studied for various problems with data coming from single source, it is not trivial for a person re-identification scenario where multi-sensor data is considered. It is a natural challenge to select a few informative samples yet cover as much appearance variation as possible across multiple cameras in such a scenario. Apart from high cost of labeling the training data, all the data may not also be available at the very outset. A static pre-trained model can not adapt to the changing dynamics of the incoming data. In this work, we focus on the fundamental challenges that need to be overcome in order to select a manageable set of training images for annotation from multi-sensor data in an online person re-identification scenario.

For this purpose, we propose an iterative framework which, starting with a pool of unlabeled images *progressively* selects the most informative yet non-redundant images - termed as the '*representative*' images. Ideally, a set of representative images are "representatives" of a dataset because this set possesses most of the variabilities of the dataset within itself. On the other hand, without any label information, the representative images are some of the most confusing samples in the whole dataset by the same trait. Thus annotating such representatives enriches the model by injecting valuable information with a reasonable labeling effort. However, in an online setting this strategy can be effective in reducing annotation effort if there is no redundancy in the selected representatives. Images captured by multiple sensors of the same persons give rise to redundant samples. Identi-

fying and eliminating redundant samples is especially important in such an active learning scenario since reducing redundancy implies more information gain at the cost of less labeling effort. The proposed work addresses the following question: *Is it possible to select a sparse set of non-redundant training images progressively in an online setting for annotation from multi-sensor data while maintaining good re-identification performance?*

Redundancy of active samples is of two types. Firstly, in each iteration, the chosen representatives may have many images of the same person. Secondly, representatives selected in subsequent iterations may also have overlap with the representatives chosen earlier for labeling. The first type of redundancy is termed as the 'intra-iteration redundancy' while the second type is termed as the 'inter-iteration redundancy'. 'intra-iteration redundancy' is restricted by exploiting the fact that redundant samples in any iteration are very close neighbors in the feature space. Without any feedback about the already chosen representatives, any representative selection strategy will tend to select images of the same persons as representatives in subsequent iterations. Though using a similar redundancy reduction strategy as above will be able to filter out samples redundant to the already labeled ones, the information gain by the system will be very little as images from new unlabeled persons will be hard to come by. We tackle this situation by enforcing diversity among the selected representatives as information about the already chosen samples in previous iterations are fed back while choosing subsequent samples to be labeled. Such a representative selection problem is formulated as a *convex optimization* that minimizes the cost of representing an unlabeled pool with a sparse set of representatives as well as one that minimizes the redundancy with the representatives selected earlier. Experiments on three benchmark datasets show that annotating the small but informative set of representative images reduces the labeling effort considerably, maintaining state-of-the-art re-identification performance.

Apart from the huge labeling effort, another factor that is a challenge for a scalable solution of the problem is the generally exponential increase of training time with the number of training samples for traditional discriminative classifiers (*e.g.*, SVM or random forest). These classifiers have to be retrained from scratch after each batch of representative selection and annotation in such repetitive active learning strategy. The generally super linear time complexity of the traditional discriminative classifiers makes them unsuitable for use in such a scenario. Though incremental learning based classifiers [92] can update the model without retraining from scratch, their performance is limited by the condition of knowing the number of classes from the start.

Motivated by the recent progress of sparse coding based classifiers [24, 113], we employ a structure preserving sparse dictionary for classification. Such a classification strategy is helpful as updating the model with newly labeled data means simply adding the new samples with labels without making any changes to the existing dictionary elements made of the already labeled samples. This model update strategy not only helps in reducing the training time significantly by avoiding the need for retraining but also enables the operation of the framework without assuming any knowledge of the number of classes. Thus, in summary, the proposed framework uses *two convex optimization based strategies to select a few informative but non-redundant samples for labeling and to update a person re-identification model online.*

The rest of the chapter is organized as follows. Section 3.2 briefly discusses the related works. An overview of the proposed approach is given in Section 3.3. Along with that the notations are also introduced. The details about the re-identification approach, as non-redundant representative selection, and the use of structure preserving sparse coding based classification are described in Section 3.4. Experimental results and comparisons are shown in Section 3.5. Finally, conclusions are drawn in Section 3.6.

3.2 Related Works

In the last few years there has been increasing attention in the fields of person re-identification, active learning and representative selection. Since the previous two chapters have detailed discussions on some of the related works in person re-identification, here we will discuss that in very brief. We will also discuss about some of the previous works in the fields of active learning and representative selection.

Person Re-identification: Person re-identification approaches are mostly supervised where the training data is processed in batch method assuming all labeled data is available beforehand. In one class of approaches [7, 59], camera invariant discriminative signatures have been used to re-identify people in different cameras. Another class [70, 91] has tried to improve the distance measure to better discriminate between different persons using simple features in a metric learning framework [26] where a non-Euclidean is learned which minimizes the distance between pairs of true matches as well as maximizes the same between pairs of wrong matches. Recently, deep convolutional architecture have enabled person re-identification to be addressed as a joint learning of discriminant signatures as well as the corresponding metric providing competitive performance [2]. However, similar to many other deep architecture, generating huge amount of labeled training data is an issue which has been addressed in the proposed work. A third class of works tried to explore transformation of features between cameras by learning brightness transfer function [48] between appearance features or different variants of it [21, 79, 93, 94]. Apart from these supervised person re-identification strategies, there has been some recent unsupervised methods [72, 118] which tried to explore saliency information or weighted features towards re-identifying people across cameras. However, none of these methods consider an interactive framework that selects the most informative set of representatives for man-

ual labeling, thus reducing the effort of the human. For a thorough review of the person re-identification literature, interested readers are directed to the review paper [109] where a multidimensional taxonomy and categorization of the person re-identification algorithms can be obtained.

Active Learning: In an effort to bypass tedious labeling of training data there has been recent interest in 'active learning' [50, 110] where classifiers are trained interactively. Queries are selected for labeling such that enough training samples are procured in minimal effort. This can be achieved by choosing one sample at a time by maximizing the value of information [50] reducing the expected error [4] or maximizing both informativeness and representativeness for active sample selection [47] prior to retraining a classifier. On the other hand there have been recent approaches [15, 30] where batches of unlabeled data are selected by exploiting classifier feedback to maximize informativeness and sample diversity. For a detailed discussion on active learning literature, the interested readers are directed to the excellent article by Settles [102].

Representative selection: Most of the applications of representative selection can be found in the fields of video summarization and subset selection. Historically clustering and vector quantization based methods [23, 37, 39] have dominated these problems, until recently sparse subset selection came into picture. In [19, 28, 29], representative selection has been formulated as sparsity regularized linear reconstruction error minimization problem. The last two works resemble most closely the proposed representative selection framework. However, without any redundancy restricting condition these frameworks can be limited in a multi-sensor application like person re-identification as far as reduction of labeling effort is concerned. A multi-sensor data has its own challenges and redundancy of representatives play a very important role in it. The Sparse Modeling Representative Selection (SMRS) framework [29] removes redundant frames from an event based summary of videos by con-

sidering the proximity of the chosen representative frames in the timeline. Time information is either unavailable in person re-identification or is unreliable for a re-identification scenario over a wide space time horizon. The proposed framework takes care of this issue by splitting the source of redundancy into two parts - one 'intra-iteration' and the other 'inter-iteration'. The 'intra-iteration' redundancy is reduced by creating a hypergraph between the chosen representatives. The redundancy among samples chosen in different iterations is reduced by introducing a convex regularization term that minimizes correlation between the new and the previously selected representatives, but at the same time chooses a number of samples representing the data aptly. This enables the selection of as many difficult examples as possible to improve the re-identification performance but at the same time avoids labeling a person multiple times unless it is necessary.

3.3 Overview of proposed approach

The overall scheme of the proposed person re-identification process is shown in Fig. 3.1. Given incoming streaming videos and detected person images, the proposed framework iteratively choses small sets of informative images to be labeled by human annotators. These informative images, called the 'active samples' are chosen starting with completely unlabeled pool of detections. At each iteration new sets of active samples are chosen from the unlabeled pool. It should be noted that the unlabeled pool can continuously get new detections from the incoming video streams. The active samples are chosen by minimizing a convex error function where the whole unlabeled pool is represented by the small set of informative samples only.

Next, the redundant images from the chosen representatives are eliminated by forming a hypergraph between the representative samples and choosing one image per hy-

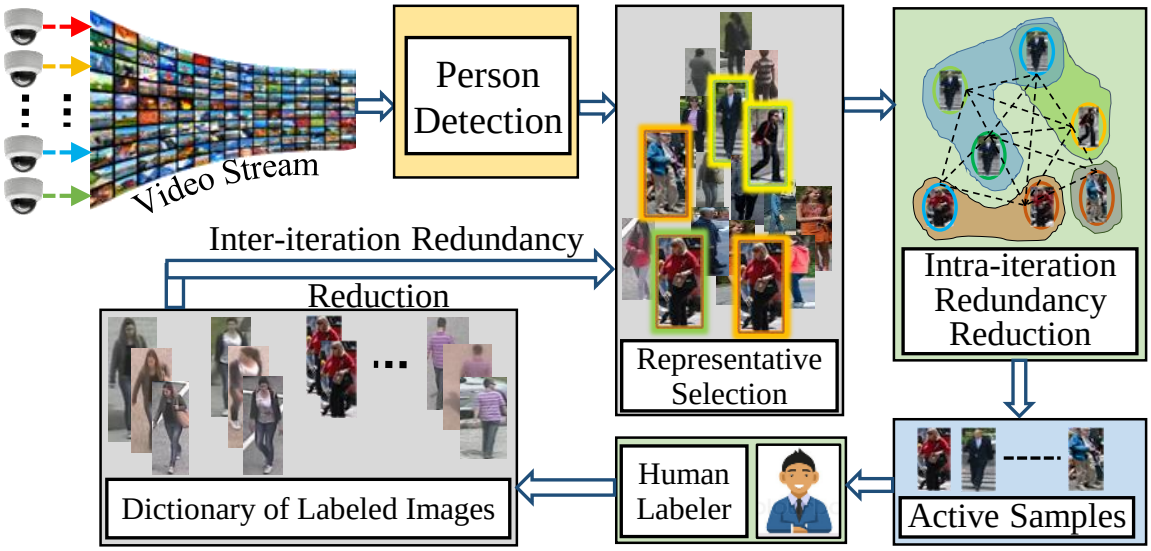


Figure 3.1: System Overview. The 'Representative Selection' module takes unlabeled images of persons and selects a few informative representatives from them. Next redundant images are removed by forming a hypergraph between the chosen samples and choosing one representative images per hyperedge. Now the active samples or the representatives filtered by the intra-iteration redundancy reduction module are presented to the human annotators seeking for labels. The labeled samples forms a dictionary which is fed to the representative selection framework so that in the next iteration those representatives from the unlabeled pool are chosen which are maximally non-redundant with the labeled images in the dictionary. This cycle goes on as new images come from the streaming videos.

peredge of the hypergraph. Overlapping hyperedges in such a hypergraph, contain images of very similar looking persons. So images which are mutually exclusive as well as the images common between the overlapping hyperedges are of utmost importance as labeling them helps in disambiguation between difficult (similar looking) persons. Now the active samples filtered by the intra-iteration redundancy reduction are presented to the human annotators for labels. These labeled samples are stacked in a dictionary which has two usages. As shown in Fig. 2.4, these images are fed to the representative selection framework along with the rest of the unlabeled images in the next iteration. The resulting convex optimization, now, minimizes the correlation between the already labeled samples and the unlabeled samples along with the reconstruction error term. This cycle goes on until a predefined number of samples are annotated. The second usage of the labeled dictionary (not shown

in the figure) is for re-identification of unknown samples in a Sparse Reconstruction based Classification (SRC) framework where the set of labeled samples work as the dictionary for the SRC.

3.4 Methodology

In this section, our proposed framework is discussed in details. First we describe the notation that would be used throughout the rest of the section before providing the problem statement, formulation and the optimization strategy to solve the problem.

3.4.1 Problem Statement

Notations Used: We use boldface uppercase letters (*e.g.*, $\mathbf{X}$) to denote matrices. A superscript(*e.g.*, $\mathbf{x}^{(i)}$)/subscript(*e.g.*, $\mathbf{x}_i$) associated with a boldface lowercase letter will denote the corresponding row/column of the matrix. A boldface lowercase letter will denote a column vector, unless otherwise specified. The ij^{th} element of the matrix $\mathbf{X}$ will be denoted as $\mathbf{X}_{ij}$. $tr(.)$ denotes the trace operator. $diag(.)$ denotes the diagonal operator which extracts the main diagonal of a matrix.

We start with a large pool of unlabeled images containing instances of persons from different cameras. This defines the input to the framework. Let at a certain iteration, the features from n unlabeled images be arranged as columns of the matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]$, where $\mathbf{z}_i \in \mathbb{R}^d$ denotes the d dimensional feature vector from the i^{th} image. We aim to select a sparse set of columns (say, k number of columns where $k \ll n$) which represents the whole collection $\mathbf{Z}$. The corresponding images are the output of the non-redundant representative selection framework which is labeled by the human annotators.

3.4.2 Basic Formulation

Finding compact dictionaries [1, 66, 119] has been studied as a way to represent data. Such approaches find the dictionary elements by searching for a set of basis vectors which expresses the data in terms of sufficiently sparse coefficient vectors with respect to the basis vectors. However, the basis vectors *i.e.*, the elements of the dictionary hardly coincides with the original data and thus do not serve as representatives selected from the data itself. To find representatives from the data itself we use the following basic formulation which defines a combinatorial optimization. This is subsequently relaxed later after it is suitably constrained by the redundancy restriction term.

$$\min_{\mathbf{D},\mathbf{U}} ||\mathbf{Z} - \mathbf{ZDU}||_F^2 \tag{3.1}$$

$$\text{s.t. } \mathbf{D} \text{ is } n \times n \text{ diagonal matrix with } diag(\mathbf{D}) \in \{0,1\}^{n\times 1},$$

$$||diag(\mathbf{D})||_0 = k, \text{ and } \mathbf{U} \in \mathbb{R}^{n\times n}$$

Here, $||.||_F$ denotes the Frobenius norm of a matrix and $||.||_0$ denotes the zero norm of a vector. $\mathbf{D}$ is a $n \times n$ diagonal matrix with only 0 or 1 in its diagonal. The ℓ_0 norm of the diagonal of such a matrix being k signifies that only k of the n diagonal elements are 1, rest are 0. Such a binary diagonal matrix when multiplied with $\mathbf{Z}$, selects only k columns out of the n columns of $\mathbf{Z}$. $\mathbf{U}$ is a full real matrix with n rows and n columns. While post multiplication of $\mathbf{Z}$ by $\mathbf{D}$ selects k columns of $\mathbf{Z}$, further post multiplication of the product by the full real matrix $\mathbf{U}$ linearly combines the selected k columns of $\mathbf{Z}$ so that the resultant matrix $\mathbf{ZDU}$ is as close as possible to the original matrix $\mathbf{Z}$. Here, both $\mathbf{D}$ and $\mathbf{U}$ are unknown. Lets denote the product of the two unknowns $\mathbf{DU}$ as $\mathbf{X}$. Since $\mathbf{D}$ is a diagonal matrix with only k 1's and $n-k$ 0's in its diagonal and $\mathbf{U}$, in general, is a full matrix, the product $\mathbf{X}$ will be matrix whose k rows will be non-zero while $n-k$ rows will be all zeros. The indices of the non-zero rows of $\mathbf{X}$ correspond to the indices of 1's in

D which, in turn, correspond to the selected columns (as representatives) of **Z**. All these characteristics of such a matrix **X** can be conveniently and succinctly expressed in terms of $\ell_{2,0}$ matrix norm. $\ell_{2,0}$ matrix norm is the number of non-zero rows of a matrix. So all the constraints in eqn. (3.1) can be written as $||\mathbf{X}||_{2,0} = k$ where $\mathbf{X} \in \mathbb{R}^{n \times n}$. Thus, changing the constraints in terms of $\ell_{2,0}$ norm, we write the basic formulation in eqn. (3.1) as follows,

$$\min_{\mathbf{X}} ||\mathbf{Z} - \mathbf{ZX}||_F^2 \tag{3.2}$$

$$\text{s.t. } \mathbf{X} \in \mathbb{R}^{n \times n}, ||\mathbf{X}||_{2,0} = k$$

The indices of the non-zero rows of **X** will give the column indices of the selected representatives from **Z**.

3.4.3 Reduction of Inter-iteration Redundancy

The above formulation selects a sparse set of representative images for labeling, but it is less effective in dealing with the 'inter-iteration' redundancy. Let us denote the set of selected representatives till a certain iteration by $\widehat{\mathbf{Z}}_0$ which is a matrix of dimension $d \times n_0$ containing the features from the already selected n_0 images. Now, **Z** contains the features from the rest of the unlabeled images. For convenience let us write the reconstructed features from this rest of the unlabeled images **ZX** as $\widehat{\mathbf{Z}}$. Without any loss of generality, let us assume that all the features are made zero mean. In that case, we show below that $||\widehat{\mathbf{Z}}_0^T \widehat{\mathbf{Z}}||_F^2$ expresses the correlation between the already selected images and the rest. This is because,

$$\begin{aligned} ||\widehat{\mathbf{Z}}_0^T \widehat{\mathbf{Z}}||_F^2 &= \sum_{i=1}^{n_0} \sum_{j=1}^{n-n_0} \left[\left(\widehat{\mathbf{Z}}_0^T \widehat{\mathbf{Z}} \right)_{ij} \right]^2 = \sum_{i=1}^{n_0} \sum_{j=1}^{n-n_0} \left[(\widehat{\mathbf{z}}_0)_i^T \widehat{\mathbf{z}}_j \right]^2 \\ &= \sum_{i=1}^{n_0} \sum_{j=1}^{n-n_0} d^4 \sigma_i^2 \sigma_j^2 \rho_{ij}^2 \end{aligned} \tag{3.3}$$

where σ_i denotes the standard deviation of the features of the i^{th} image in $\widehat{\mathbf{Z}}_0$ and likewise σ_j denotes the standard deviation for the j^{th} image in $\widehat{\mathbf{Z}}$. ρ_{ij} denotes the correlation coefficient between the features of the i^{th} image in $\widehat{\mathbf{Z}}_0$ and the j^{th} image in $\widehat{\mathbf{Z}}$. The last line in eqn. (3.3) is due to the fact that all the columns of both $\widehat{\mathbf{Z}}_0$ and $\widehat{\mathbf{Z}}$ have been converted to zero means. From this, it can be seen that minimizing $||\widehat{\mathbf{Z}}_0^T\widehat{\mathbf{Z}}||_F^2$ prefers to select the columns of $\mathbf{Z}$ which are less correlated to the images in $\widehat{\mathbf{Z}}_0$. So, adding $||\widehat{\mathbf{Z}}_0^T\widehat{\mathbf{Z}}||_F^2$ (*i.e.*, $||\widehat{\mathbf{Z}}_0^T\mathbf{ZX}||_F^2$) as a regularizer to eqn. (3.2) makes sure that a sparse set of images non-redundant with previously selected representatives are obtained. Using a regularization parameter λ_1 the problem can now be written as,

$$\min_{\mathbf{X}} ||\mathbf{Z} - \mathbf{ZX}||_F^2 + \lambda_1||\widehat{\mathbf{Z}}_0^T\mathbf{ZX}||_F^2 \tag{3.4}$$

$$\text{s.t. } \mathbf{X} \in \mathbb{R}^{n\times n}, ||\mathbf{X}||_{2,0} = k$$

In Eqn. (3.4), the first term of the cost function minimizes the reconstruction error of the feature from each image when the reconstruction is done as a linear combination of features from the selected representative images. The second term minimizes the correlation between the selected representatives with the previously selected ones. The constraint on $\ell_{2,0}$ norm of $||\mathbf{X}||$ implies that only k rows of it will be non-zero. In the reconstruction term, $\mathbf{ZX}$, x_{ij} is multiplied with $\mathbf{z}_i$ towards the reconstruction of the j^{th} column of $\mathbf{Z}$. Thus, if $\mathbf{x}^{(i)}$ contains all zeros (*i.e.*, $x_{ij} = 0, \forall j$), that means $\mathbf{z}_i$ is not contributing anything towards the reconstruction of any column of $\mathbf{Z}$. Thus, $\mathbf{z}_i$ is not a good representative of $\mathbf{Z}$. As a result, the nonzero rows of $\mathbf{X}$ correspond to those columns of $\mathbf{Z}$ which represent the whole unlabeled pool $\mathbf{Z}$.

3.4.4 Relaxation of the Constraints

Eqn. (3.4) is NP-hard and can be highly computationally expensive even for moderate values of k and n. We need to relax the optimization problem in eqn. (3.4) to make it a convex optimization problem as the $\ell_{2,0}$ norm is non-convex. Following the common strategy of 1-norm relaxation for 0-norms, we employ $||.||_{2,1}$ norm in place of the $\ell_{2,0}$ norm and reformulate the problem as,

$$\begin{gathered} \min_{\mathbf{X}\in\mathbb{R}^{n\times n}} ||\mathbf{Z}-\mathbf{Z}\mathbf{X}||_F^2 + \lambda_1||\widehat{\mathbf{Z}}_0^T\mathbf{Z}\mathbf{X}||_F^2 \\ \text{s.t. } ||\mathbf{X}||_{2,1} \leq k \end{gathered} \tag{3.5}$$

3.4.5 Overall Optimization Problem

Using Lagrange multipliers, the overall optimization problem from eqn. (3.5) can be written as,

$$\min_{\mathbf{X}} ||\mathbf{Z}-\mathbf{Z}\mathbf{X}||_F^2 + \lambda_1||\widehat{\mathbf{Z}}_0^T\mathbf{Z}\mathbf{X}||_F^2 + \lambda_2||\mathbf{X}||_{2,1} \tag{3.6}$$

where, λ_1 and λ_2 are the two regularization parameters. The inputs to the optimization problem are the unlabeled images $\mathbf{Z}$ and the labeled images $\widehat{\mathbf{Z}}_0$ while the output is the selection matrix $\mathbf{X}$ whose non-zero row indices provide the representative images to be labeled. After labeling, the annotated samples are inducted into the dictionary as dictionary elements and the re-identification probability of the test images are obtained by finding sparse representations of the test samples with respect to the dictionary according to the formulation described next.

The conversion of the constrained optimization problem to the corresponding unconstrained problem as in eqn. (3.6), employing Lagrange multipliers brings in independence from k. That is, the number of non-zero rows may not be exactly k. Following standard practice in literature [19, 29] we choose the top k rows in terms of their 2 norms when the number of non-zero rows of $\mathbf{X}$ is greater than k. For the case when this number is

less than k, only all the non-zero rows are taken. The representatives are chosen from the corresponding columns of $\mathbf{Z}$.

3.4.6 Reduction of Intra-iteration Redundancy

We have seen that $\mathbf{ZX}$ gives the reconstructed pool of unlabeled images as a linear combination of the selected representatives where the selected representatives of $\mathbf{Z}$ are given by the indices of the non-zero rows of $\mathbf{X}$. Due to the presence of real numbers in the product $\mathbf{X}$ there can be repetitive selection of images resulting in intra-iteration redundancy between the selected representatives. This is reduced by forming a hypergraph among the representatives selected by solving eqn. (3.5). hypergraphs [12] are a generalization of graphs where one edge can be connected to any number of edges. Such an edge is named as a hyperedge which links a subset of nodes instead of two nodes only in ordinary graphs. In this sense, an ordinary graph is a special kind of hypergraph. After each iteration, such a hypergraph is formulated where the nodes of the hypergraph are the chosen active samples in that particular iteration. The hyperedges, created in the feature space itself, contains the redundant samples. From the '$k \times k$' feature similarity matrix, a '$k \times k$' *adjacency matrix* is created using a high threshold of feature similarity. The adjacency matrix subsequently gives the'$m \times k$' *incidence matrix* where 'm' is the number of hyperedges. Note that such a graph based clustering has major advantage over popular and simple clustering methods e.g., k-means as the success of k-means depends largely on the judicious choice of k. While hypergraph based redundancy reduction depends on the threshold of the similarity score, we set it very high as a high threshold of similarity scores makes sure that only very similar samples qualify as redundant samples. As all the images in each group of redundant samples are given a single identity, the use of such high threshold prevents the model to get updated with wrong labels.

3.4.7 Classification and Online Update

The chosen samples are annotated by the human annotators and the annotated samples form the dictionary elements. The dictionary is used to find the probability of the test samples via finding the sparse representations of the test samples. Using the annotated representatives $\widehat{\mathbf{Z}}_0$, as a dictionary, the sparse representation of the test samples $\mathbf{Y}$ can be found by minimizing the following.

$$\min_{\mathbf{C}} ||\mathbf{Y} - \widehat{\mathbf{Z}}_0\mathbf{C}||_F^2 + \alpha||\mathbf{C}||_1 \tag{3.7}$$

Ideally a test image is reconstructed from a linear combination of labled samples from the same class as that of the test sample. The sparsity condition makes sure that training samples from other classes appear as infrequently as possible in the reconstruction of the test image. Seeking the sparsest representation, therefore, discriminates between the various classes of test samples and the sparse coefficients (when normalized) provide the probability of the test sample to belong to that class. However, the overcomplete nature of the dictionary can give rise to loss in structure of the data. Similar features may be encoded by different sparse codes giving rise to entirely different probability distribution for samples of same class [96].

To increase the robustness of a sparse code based classifier, graph Laplacian has been used [38, 119]. After incorporating the structure preserving regularizer in eqn. (3.7), the sparse classifier can be written as,

$$\min_{\mathbf{C}} ||\mathbf{Y} - \widehat{\mathbf{Z}}_0\mathbf{C}||_F^2 + \alpha||\mathbf{C}||_1 + \beta\, tr(\mathbf{C}\mathbf{L}\mathbf{C}^T) \tag{3.8}$$

where $\mathbf{L}$ is the graph Laplacian [76] obtained from a k-nearest neighbor graph of similarities calculated between the columns of $\mathbf{Y}$. Using such a sparsity based strategy we are able to update the classifier online simply by incorporating the labeled images in any iteration to the

already existing dictionary. Unlike the discriminative classifiers this involves no expensive training phase and thus online update of the classification model is not an overhead with large number of classes.

3.4.8 Optimization

Here we state the optimization strategy to solve the two convex optimization problems (eqns. (3.6) and (3.8)). Both the equations involve convex but non-smooth terms which require special attention. Proximal methods are specifically tailored towards it. These methods have drawn increasing attention in the machine learning community because of their fast convergence rates. They find the minimum of a cost function of the form $g(\mathbf{X}) + h(\mathbf{X})$ where g is convex, differentiable but h is closed, convex and non-smooth. We use fast proximal algorithm, FISTA [9] which maintains two variables in each iteration and combines them to find the solution. New value of the variable, in each iteration is computed by computing the proximal operator of h on a function of the gradient of g. (ref eqn. (3.10)). The proximal operator of $h(\mathbf{X})$, denoted as $\text{Prox}_h(\mathbf{X})$ is computed as,

$$\text{Prox}_h(\mathbf{X}) = \underset{\mathbf{U}}{\text{argmin}} \left(h(\mathbf{U}) + \frac{1}{2}||\mathbf{U} - \mathbf{X}||^2 \right) \tag{3.9}$$

where, the domain of $\mathbf{U}$ is the set of real matrices with same dimension as $\mathbf{X}$. The FISTA algorithm can be summarized by the following two steps after choosing any initial $\mathbf{X}^{(0)} = \mathbf{X}^{(-1)}$ (the superscripts, here, denote the iteration number of FISTA).

$$\begin{aligned} \textbf{Step I: } \mathbf{Y} &= \mathbf{X}^{(k-1)} + \frac{k-2}{k-1}\left(\mathbf{X}^{(k-1)} - \mathbf{X}^{(k-2)}\right) \\ \textbf{Step II: } \mathbf{X}^{(k)} &= \text{Prox}_{t_k h}(\mathbf{Y} - t_k \nabla g(\mathbf{Y})) \end{aligned} \tag{3.10}$$

where, k is the iteration index (of FISTA) and t_k is the step size parameter.

Gradients and Lipscschitz's constants: In eqn. (3.6), the reconstruction error and the inter-iteration redundancy reduction terms are convex, smooth, differentiable functions with Lipschitz continuous gradients. Let us denote the sum of these two terms as $g(\mathbf{X})$ *i.e.*,

$$g(\mathbf{X}) = ||\mathbf{Z} - \mathbf{ZX}||_F^2 + \lambda_1 ||\widehat{\mathbf{Z}}_0^T \mathbf{ZX}||_F^2 \tag{3.11}$$

The gradient $\nabla g(\mathbf{X})$ and the Lipschitz constant L_g of the gradient are given by,

$$\begin{aligned} \nabla g(\mathbf{X}) &= 2\big(- \mathbf{Z}^T\mathbf{Z} + \mathbf{Z}^T\mathbf{ZX} + \lambda_1 \mathbf{Z}^T \widehat{\mathbf{Z}}_0 \widehat{\mathbf{Z}}_0^T \mathbf{ZX}\big) \\ L_g &= 2\big(||\mathbf{Z}^T\mathbf{Z}||_F^2 + \lambda_1 ||\mathbf{Z}^T \widehat{\mathbf{Z}}_0 \widehat{\mathbf{Z}}_0^T \mathbf{Z}||_F^2\big) \end{aligned} \tag{3.12}$$

Similarly, the reconstruction error and the structure preserving terms in eqn. (3.8), are convex, smooth and differentiable functions of $\mathbf{C}$. Denoting $||\mathbf{Y} - \widehat{\mathbf{Z}}_0\mathbf{C}||_F^2 + \beta tr(\mathbf{CLC}^T)$ as $p(\mathbf{C})$, the gradient $\nabla p(\mathbf{C})$ and the Lipschitz constant L_p of the gradient are given by,

$$\begin{aligned} \nabla p(\mathbf{C}) &= 2\big(- \widehat{\mathbf{Z}}_0^T\mathbf{Y} + \widehat{\mathbf{Z}}_0^T \widehat{\mathbf{Z}}_0 \mathbf{C} + \beta \mathbf{CL}\big) \\ L_p &= 2\big(||\widehat{\mathbf{Z}}_0^T \widehat{\mathbf{Z}}_0||_F^2 + \beta ||\mathbf{L}||_F^2\big) \end{aligned} \tag{3.13}$$

Proximal operators: The sparsity inducing $\ell_{2,1}$ norm (in eqn. 3.6) and the ℓ_1 norm (in eqn. 3.8) both are convex but non-smooth functions of their respective variables. Let us denote the non-smooth terms as $h(\mathbf{X})$ and $q(\mathbf{C})$ respectively, *i.e.*, $\lambda_2||\mathbf{X}||_{2,1} = h(\mathbf{X})$ and $\alpha||\mathbf{C}||_1 = q(\mathbf{C})$. The corresponding proximal operators for these two non-smooth functions are given by,

$$\text{Prox}_h(\mathbf{X}) = \big(1 - \frac{\lambda_2}{||\mathbf{X}^{(i)}||_2}\big)_+ \mathbf{X}^{(i)} \tag{3.14}$$

$$\text{Prox}_q(\mathbf{C}) = \big(1 - \frac{\alpha}{|\mathbf{C}_{ij}|}\big)_+ \mathbf{C}_{ij} \tag{3.15}$$

where i and j denote the row and column numbers with $(x)_+ \triangleq \max(x, 0)$. Taking a fixed step size t_k equal to the inverse of the respective Lipschitz constants, the convergence rate of FISTA is proportional to $\frac{1}{k^2}$, in contrast to $\frac{1}{\sqrt{k}}$ in subgradient based methods where k denotes the iteration number. The overall iterative framework towards online and interactive

person re-identification using the gradients, Lipschitz constants and the proximal operators is presented in algorithm 1.

Algorithm 1 Overall Framework

Active Training:

Input: Unlabeled images $\mathbf{Z}, \lambda_1, \lambda_2, T$ (# of iterations)

Output: Representatives for labeling $\widehat{\mathbf{Z}}_0$

$\widehat{\mathbf{Z}}_0 \leftarrow \phi$ (null set),

for $i \leftarrow 1$ to T **do**

$\mathbf{X} \leftarrow$ solution of eqn. (3.6) by FISTA (eqn. 3.10 and 3.14) using gradient $\nabla g(\mathbf{X})$ and Lipschitz constant L_g (eqn. 3.12)

$\mathbf{Z}_s \leftarrow$ columns of $\mathbf{Z}$ corresponding to non-zero rows of $\mathbf{X}$

$\widehat{\mathbf{Z}}_0 \leftarrow \widehat{\mathbf{Z}}_0 \cup \mathbf{Z}_s$, $\mathbf{Z} \leftarrow \mathbf{Z} \setminus \mathbf{Z}_s$

end for

Test:

Input: $\mathbf{Y}, \widehat{\mathbf{Z}}_0, \mathbf{L}, \alpha, \beta$

Output: Sparse coefficient matrix C

$\mathbf{C} \leftarrow$ solution of eqn. (3.8) by FISTA (eqn. 3.10 and 3.15) using gradient $\nabla p(\mathbf{C})$ and Lipschitz constant L_p (eqn. 3.13)

3.5 Experiments

The experiments are designed keeping the following three main objectives in mind. **Objective 1:** First of all, we will analyze how the proposed framework helps in getting better re-identification performance (in terms of re-identification accuracy) by choosing a sparse set of informative samples for annotation. For this purpose we will compare

the re-identification accuracy vs the number of images labeled, with the following three baselines. **Baseline-I** assumes that the representatives are chosen randomly for labeling *but* a discriminative classifier - linear SVM is used for classification instead of the sparse code based classifier. Though SVMs with any non-linear kernel or other nonlinear classifiers *e.g.*, a Random Forest would have equally served the purpose, we chose linear SVM over those as it requires less training time and the training is independent of any tunable parameter. **Baseline-II** - In addition to using linear SVM as the classifier, this baseline chooses the images following the proposed framework (Sec. 3.4.5) for labeling. **Baseline-III** - This baseline chooses samples randomly for annotation while the classifier used here is SRC.

We also compared with a state-of-the-art representative selection framework - Sparse Modeling Representative Selection (SMRS) [29] which does not consider redundancy among chosen representatives. While comparing with the baselines shows the significance of informative representative selection over random selection for active labeling, the comparison with SMRS shows the role of redundancy reduction in the online setting. For this purpose, we conducted experiments starting with unlabeled images with both balanced and imbalanced distributions of images per person. Balanced and imbalanced scenarios are described in detail in section 3.5.1.

Objective 2: The next objective is to study the scalability of the approach with a dataset containing a large number of persons. The dataset considered here is an order of magnitude larger than in the above objective with respect to the number of people. The performance measures and comparison baselines for this case are the same as in Objective 1.

Objective 3: Last of all, we also compare with other state-of-the-art methods with the help of performance measures traditionally used (Cumulative Matching Characteristic (CMC)) in supervised person re-identification methods.

There are many datasets that can be used to evaluate the proposed method for the said objectives. However, some of these datasets (*e.g.*, VIPER, GRID) contain too few images per person to form disjoint train and test sets in an online person re-identification scenario. As a result, we chose to experiment with three benchmark datasets - WARD [80], iLIDS-VID [112], and CAVIAR4REID [17] - which not only contains a lot of persons but also has several repetitive images per person giving the opportunity to show performance improvement in an incremental manner as more and more unlabeled data are annotated.

Feature Extraction: Mean color feature (HSV) is used following the scheme in [45]. Since the images are from different cameras, the features can vary a lot due to the changes of several factors including but not limited to scale, illumination, depth *etc.* However, for a single person as the features are coming from the same person irrespective of the camera, it is reasonable to assume that the features from the same person are close to each other in some underlying joint manifold. This directed us to find a low dimensional manifold out of the features from the unlabeled pool of images. Considering the success of t-SNE [25] in finding a low dimensional nonlinear embedding of unlabeled data while maintaining neighboring structure of the data in the high dimensional space, we found a low dimensional t-SNE representation of the features before proceeding further.

Experimental Setup:

- Images have been normalized to 128×64 to be consistent with the state-of-the-art person re-identification methods.
- After segmenting the images into three salient regions (head, torso and legs) [7], mean color feature (HSV) is generated following the scheme in [45]. The head region is discarded, as it consists of a few and less informative pixels. Each bodypart, is divided into blocks of size 8×16 and the blocks are overlapping by 50% in horizontal and vertical directions.

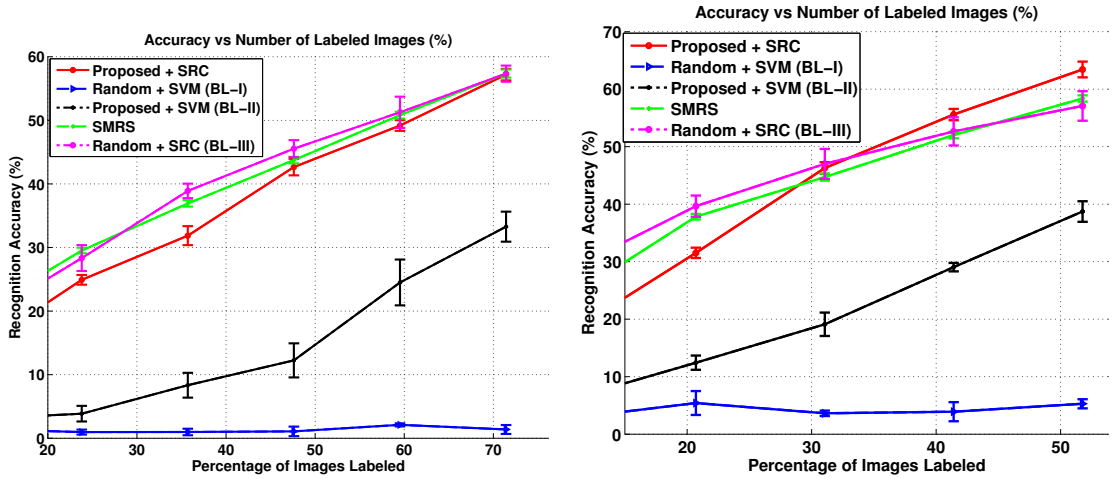
- The regularization parameter λ_2 is taken as λ_0/γ where λ_0 is computed from the data [28] and γ is taken as 2.5 throughout. For the other parameters, following values were used throughout, $\alpha = 0.2$ and $\beta = 0.3$. For both WARD and CAVIAR4REID, λ_1 is taken as 2. Since the number of people and images is more in iLIDS-VID, redundant examples are abundant compared to the other two datasets and thus λ_1 is taken as 10.
- The dimension of the joint manifold has been taken as 10 throughout.
- We ran all the experiments with 5 independent trials and report the average results. For each trial a unlabeled pool and a separate disjoint test set were created randomly.
- For both basline I and III 10% images of the starting unlabeled pools were chosen randomly for annotation in each iteration. The exact values for each dataset are given in the following subsections. For fair comparison, we chose same number of images in case of the proposed method when the number of the chosen representatives is more than this number.
- The threshold for intra-iteration redundancy reduction was taken as 0.8 in the scale of similarity scores between 0 and 1. To compare fairly, the intra-iteration redundancy reduction step was applied to random selections too (*i.e.*, in baseline II and III) with same threshold value.
- We used the toolbox LIBSVM [16] implementation for the linear SVM classifier.
- The proposed framework, generally, chooses different number of samples for labeling in each iteration for different unlabeled pool. As a result, for different test sets, the accuracy may not be obtained for the same number of images labeled. So we used spline interpolation to get the accuracies for the same number of labeled image. For each experiment, the average accuracy vs labeled images plots were calculated taking

the mean of this interpolated plots. To show the robustness of the methods, we also show the corresponding ± standard deviation values of the accuracies too.

3.5.1 WARD Dataset

The WARD dataset [80] has 4786 images of 70 different people acquired in a real surveillance scenario in 3 non-overlapping cameras. It has large illumination variation along with resolution and pose changes. This dataset is used to show the performance of the proposed framework starting with a balanced and an imbalanced pool of unlabeled images as two separate scenarios. By 'balanced' we mean that the pool is composed in such a way that each person has equal number of images per camera. Though, in reality, such a perfectly balanced distribution of data is hard to come by, we conducted the experiments on such a balanced scenario to show that the proposed method performs well in such a scenario too. For this dataset 2 random images per person per camera was chosen to form such a balanced pool. The imbalanced pool was formed such that 20% of the persons (*i.e.*, 14) have 10 images, 50% persons (*i.e.*, 35) have 4 images and 30% persons (*i.e.*, 21) have 2 images per camera. The test set for both the cases is composed of 2 images per person per camera.

Fig. 3.2(a) and (b) show the comparative analysis of the test set accuracies as a function of the number of images labeled (as a percentage of the number of starting unlabeled images). While the plots show the mean accuracy over 5 independent trials the vertical bars in each of the plot denote the corresponding standard deviation of the accuracy values around the mean. In the balanced scenario, the number of images in the unlabeled pool to start with is 420 (70*2*3) and the accuracies are shown till around 70% of the images are labeled. For the imbalanced scenario the number of images in the starting pool



(a) (b)

Figure 3.2: Plot of testset accuracy (average) with the percentage of images labeled for the WARD dataset. Fig. (a), (b) show the performances for balanced and imbalanced set of unlabeled pools respectively.

is 966 and accuracies are shown till around 50% of the images are labeled in Fig. 3.2(b). For random selections (*i.e.* for baselines I and III) in the balanced scenario, 50 random images (10% of 420 unlabeled images and then rounded to nearest greater multiple of 10) are chosen for annotation in each iteration. In the imbalanced scenario, the number is 100 (10% of 420 unlabeled images and then rounded to nearest greater multiple of 10).

Analysis of the Results: For both the balanced and the imbalanced data distribution, a discriminative classifier (baseline I and II) always underperforms compared to the cases where SRC is used as the classifier irrespective of the method of active sample selection (baseline III, SMRS and the proposed method). Apart from lower accuracies, both baseline I and II require more training time than SRC as the discriminative classifier needs to be retrained from scratch after each iteration of active sample selection. However, a comparison between the performances of baselines I and II shows that the proposed method of active sample selection provides better recognition accuracy for a fixed annotation effort.

Interesting observations can be made when performances of the three scenarios (viz. Baseline III, SMRS and proposed method) are compared where the classifier is kept fixed as SRC. For the balanced scenario, it can be seen that the three methods perform pretty closely. Though SMRS and baseline III follow each other very closely, baseline III is more uncertain than both the representative selection based methods with or without considering redundancy. This is shown by higher values of standard deviations for baseline III. The superiority of the proposed method over SMRS can be observed in the more practical scenario when the data distribution is imbalanced (Fig. 3.2(b)). Starting with lower recognition accuracy than both SMRS and baseline III, the proposed method surpasses SMRS when around 28.5% images are annotated while it surpasses baseline III when around 37.5% images are annotated. With 50% annotated data, the performance of the proposed method is better than the next best (SMRS) by around 5%.

When compared between the balanced and imbalanced scenarios, the uncertainty for all the methods are seen to be more for the imbalanced pool. The imbalance in data distribution is, thus, seen to affect all the methods but the relatively large value of the error bars for baseline III where random selection of images are made shows that imbalance affects the proposed method less than it affects random selection. This is due to the reason that, in random representative selection, the samples are selected for annotation following a similar imbalanced distribution as the original pool. On the other hand, the proposed method judiciously selects a diverse set of samples to negate the effect of imbalance in the data. This can be seen more precisely in Fig. 3.3 where the three bars represent the number of samples per person in the starting imbalanced pool (black), in the annotated sets with proposed framework (red) and random selection (green) after 25% of the unlabeled images are chosen by these two methods for labeling. The horizontal axis shows the person IDs. The distribution of images for annotation is seen to roughly follow the same distribution

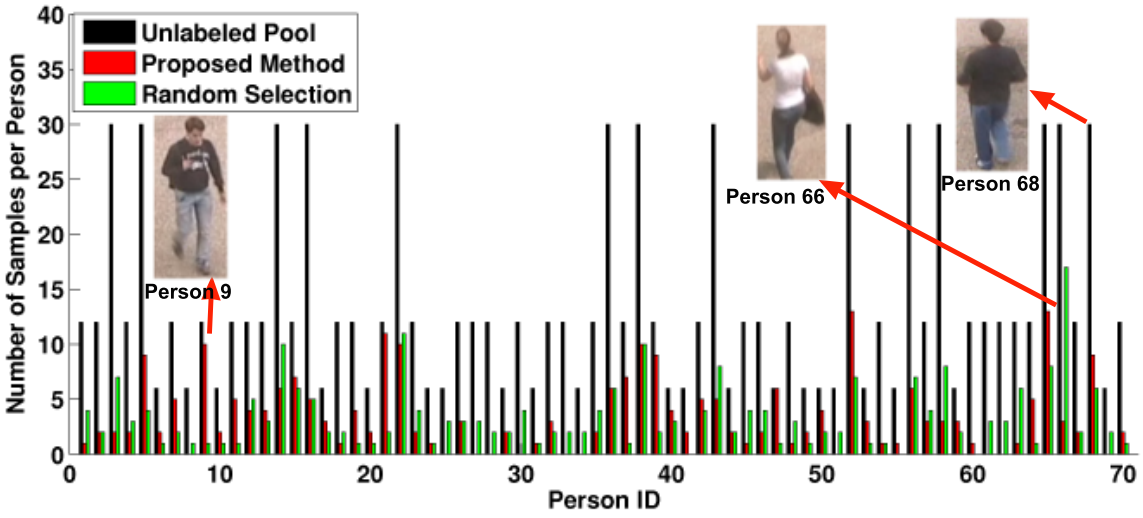
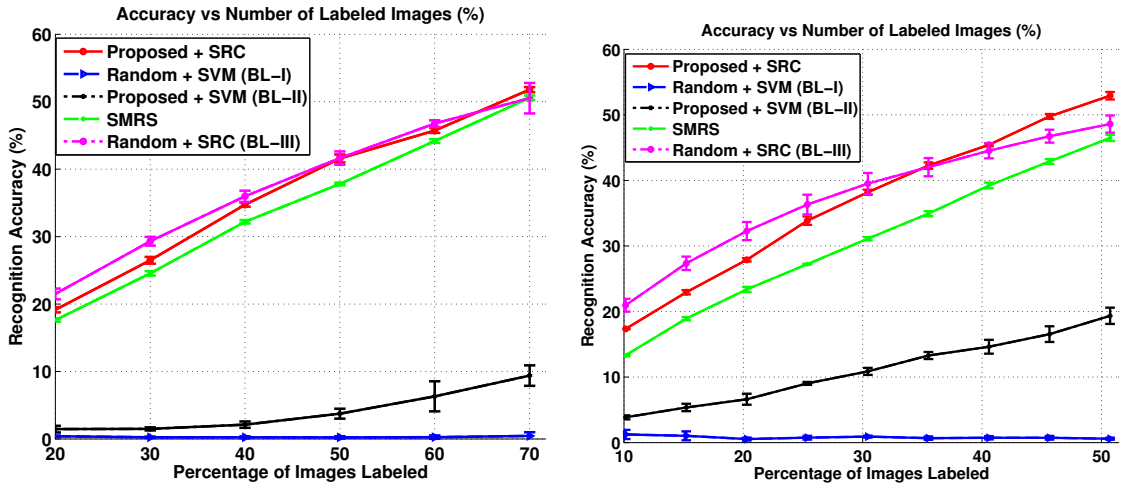


Figure 3.3: Imbalanced pool of unlabeled images. The three bars for each person (Id of the person is in the horizontal axis) give the number of images of that person in the starting unlabeled pool (black), in the annotated sets with proposed framework (red) and random selection (green). This snapshot is given after 25% of the images in the imbalanced pool are labeled for each of the methods. See text (Sec. 3.5.1) for a detailed analysis of this figure.

as that of the pool for the random selection while that is not the case for the proposed framework. For example, person 68 and 9 look very similar and the proposed method chooses more number of images for both of them as they can create confusion than say, person 66 who looks markedly different. This is done irrespective of the original distribution of the unlabeled pool.

3.5.2 i-LIDS-VID Dataset

iLIDS-VID [112] is a recently introduced person re-identification dataset. This dataset consists of images from 300 people at an airport arrival hall captured through 2 non-overlapping cameras. Apart from the typical challenges in person re-identification *e.g.*, clothing similarities, clutter, lighting variations *etc.*, one significant challenge in this dataset is the large number of people to be re-identified. Following the same convention with WARD, here also we experiment in two scenarios - balanced and imbalanced data



(a) (b)

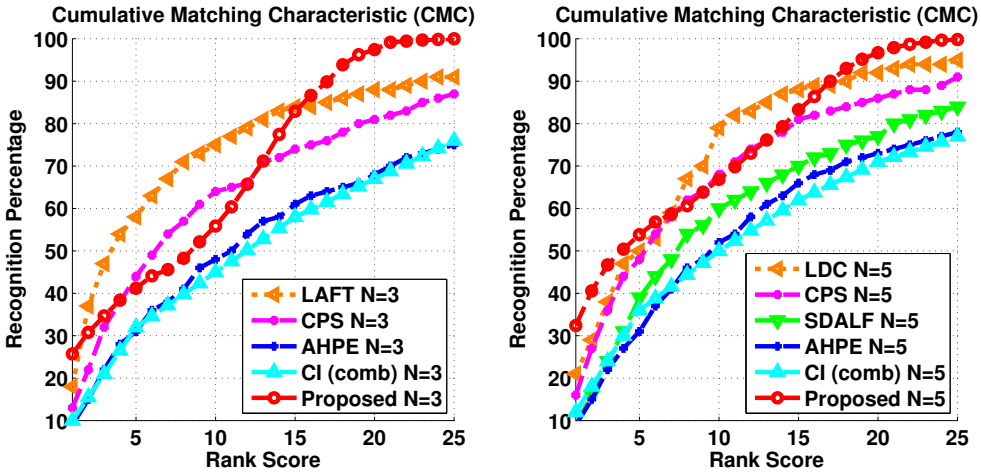
Figure 3.4: Plot of testset accuracy (average) with the number of images labeled for the i-LIDS-VID dataset. Fig. (a), (b) show the performances for balanced and imbalanced set of unlabeled pools respectively.

distributions. The composition of unlabeled pool for both the scenarios are exactly same as the WARD dataset. However, due to the presence of more number of persons, the number of unlabeled images are much more than WARD. The numbers are 1200 for the balanced scenario and 2760 for the imbalanced scenario compared to 420 and 966 respectively for WARD. The test set for this dataset also is composed of 2 images per person per camera. Fig. 3.4(a) and (b) show the comparative analysis of the test set accuracies for this dataset as a function of the number of images labeled for the balanced and imbalanced set of unlabeled pools. Keeping the same convention with WARD, the accuracies are shown till 70% of the images in the unlabeled pool are labeled for the balanced scenario while for the imbalanced scenario accuracies are shown till around 50% of the images in the unlabeled pool are labeled. For random selections (*i.e.* for Baseline I and III) in the balanced scenario, 120 random images (10% of 1200 unlabeled images) are chosen for annotation in each iteration. In the imbalanced scenario, the number is 280 (10% of 2760 and then rounded to nearest greater multiple of 10).

Analysis of the Results: For this dataset also, it can be seen that a discriminative classifier (baseline I and II) underperforms compared to the cases where SRC is used as the classifier irrespective of the method of active sample selection (baseline III, SMRS and the proposed method) for both balanced and imbalanced scenarios. For the later 3 methods also the trend is similar to that seen in WARD. When compared to the results of the WARD dataset, for both the scenarios, the accuracies with same percentage of labeled images are less for all these three methods. For example the accuracies of the proposed method, SMRS and baseline-III in the balanced scenario is 51.79%, 50.58% and 50.52% respectively for the i-LIDS-VID dataset with 70% labeled images compared to 56.25%, 56.55% and 56.60% for the WARD dataset at the same percentage of labeled images. This is due to the more variability present in the i-LIDS-VID dataset with increased number of persons. The significant fall in performance with baseline II and I shows that random selection or a discriminative classification strategy with comparatively less informative samples are less effective in presence of huge variation in the data. For the imbalanced data distribution the uncertainty is more in case of random selection (baseline-III) than both the representative selection based strategies (proposed and SMRS) with or without considering redundancy. This is seen by the larger value of the standard deviations for baseline III.

3.5.3 CAVIAR4REID Dataset

This dataset [17] contains images of pedestrians extracted from the CAVIAR repository. It is composed of 1000 images of 50 pedestrians viewed by two disjoint cameras. The challenges in this dataset involve a broad change in the image resolution from 17×39 to 72×144 with severe pose variations, illumination changes and occlusion. Many state-of-the-art approaches have evaluated their performance on this dataset. For this dataset, it



(a) (b)

Figure 3.5: CMC curves for CAVIAR4REID dataset. In (a) and (b) comparisons are shown with the state-of-the-art methods in multishot strategies with N=3 and N=5 respectively. See text for the definition of N.

is common to evaluate the performance in two different scenarios, namely the 'single shot' and the 'multiple shot' strategies. In single shot strategy both the train and test set is composed of a single image per person in each camera. Since this strategy does not involve redundancy and moreover the unlabeled pool (from which the training set will be formed) is not large in size, we don't evaluate in this scenario. In the multiple shot strategy, the train and test set is composed of $N(>1)$ images per person. As a result, the traditional approaches, in this dataset, use $50 \times N$ images for both training and test purposes. In our experiments we also create such a test set by selecting N random images per person. The rest of the images $(1000 - 50 \times N)$ form the starting unlabeled pool. To compare fairly with the traditional methods we show the CMC curve when the proposed method has labeled equal number of images as these methods. In particular we took N=3 and 5 in two different settings. For N=3, we start with an unlabeled pool of 850 images and compute the CMC curve when 150 images have been labeled (ref. Fig. 3.5(a)). Similarly, for N=5, the size of the starting unlabeled pool is 750 and the CMC is computed when the number of labeled

images reach 250 (ref. Fig 3.5(b)). We compared with the following methods - AHPE [8], SDALF [7], CI [59], CPS [17], LAFT [69] and LDC [117]. Some of these methods have published results for both the values of N while some have for only one of the two values of N.

Analysis of the Results: For both 'N=3' and 'N=5' the proposed method has the top rank 1 performance. In 'N=3' scenario, the rank 1 performance of the proposed method is 25.73% compared to the second best of 18% given by LAFT, while in 'N=5' scenario, the rank 1 performance of the proposed method is 32.4% compared to the second best of 21% given by LDC. Though for rank 8 to 16 in 'N=5' scenario, the proposed method is only second to LDC, it recovers soon and reaches the 100% recognition performance the earliest. Similar trend is seen for the 'N=3' scenario too. Reaching the 100% early is desirable in a re-identification scenario as that means the true match of the sample images are found with certainty within a small number of trials.

3.6 Conclusions

In this work, we addressed the problem of re-identifying persons in an active learning set up with two different goals - reducing the labeling effort and the training time by updating the model online. In doing so, a convex optimization based framework is proposed that progressively and judiciously chooses a sparse and non-redundant set of samples for labeling. A sparse representation based classifier is used for online updation of the model. Experiments on three publicly available benchmark datasets are performed to validate the proposed approach. The future directions of our research will be to apply the framework to bigger networks with large numbers of cameras, and cope with wider horizons of computer vision *e.g.*, online and continuous activity recognition.

Chapter 4

Attribute Based Active Learning for Continuous Person Re-identification

Continuing with the idea of involving human in boosting re-identification performance this chapter also looks to reduce costly human labor but at the same time maintain good re-identification performance. In particular we leverage upon the 'value of information' [50] active learning framework to reduce human effort especially when classification is to be performed over many categories. In this framework yes-no type binary answer instead of a precise category label is sought from the human annotators. Mid level semantic features called 'attributes' are used effectively with this framework for efficiently involving human in the loop to re-identify persons with continuous inflow of unlabeled data. Form the unlabeled pool of images, the query image (*i.e.*, the unlabeled image to be labeled) is selected such that the humans response is likely to reduce the subsequent misclassification risk the most at the cost of least annotation effort from the human. Along with labels

for the unknown persons, the proposed approach makes sure that the human also provides an explanation for the decision which is subsequently incorporated in the model to help reducing the human effort for similar kind of examples. The explanation is given in terms of a language that is understood by both man and machine. Attributes define such a richer language to convey the domain knowledge from the expert to the model. We demonstrate the effectiveness of the proposed method for continuous person re-identification system with two datasets.

4.1 Introduction

Traditional person reidentification are static and mostly supervised. In this work, we focus on the fundamental challenges that need to be overcome in order to address the largely unaddressed problem of continuous adaptation of person re-identification models starting with a small pool of labeled images. In short, we term this as 'continuous person re-identification'.

In the presence of a continuous inflow of unlabeled images containing both previously seen and unseen persons, inputs from a human is necessary. However, the human has to invest a considerable amount of effort in labeling an unlabeled image, especially in presence of a lot of visually similar persons. Thus, a scalable solution to such a problem requires a small number of questions to be asked to the human without compromising the performance. Towards this goal, the system can use feedback from the human expert so that knowledge from the human is transferred and is reflected in asking questions that are fewer in number but better in quality. This work proposes an active learning based continuous person re-identification framework which incorporates the knowledge of the human in the loop to reduce the labeling effort, while allowing the model to be continuously updated.

Traditional active learning settings involve tedious comparisons with all the classes by a human. The incorporation of the domain knowledge from the human to the process can help in reducing the subsequent effort in labeling. A recent line of work [32, 87, 88] draws inspiration from the way human experts simplify the task of discrimination by using mid level semantic features, called *attributes*. Attributes define a richer language to convey the domain knowledge from the expert to the model. Inspired by the recent success of using attributes as feedback in face recognition and scene classification [13, 32, 89], we combine attribute feedback with 'value of information' [50] based active learning strategy to select a small but informative set of images for labeling.

Though some recent works in re-identification [64, 65, 105, 73] have studied the use of attributes, they used it as a replacement to low level features. Unlike these works where pre-annotated data with a predefined vocabulary of attributes were assumed, the proposed framework uses the attribute feedback to learn attribute predictors on the way. Active learning methods with or without involving attributes, often, depend on the assumption of having training examples from all possible classes at the start. A simple way of alleviating this restriction is by setting up a threshold of maximum number of comparisons before giving a new label [50]. These assumptions are unrealistic for real life re-identification problems where new persons come in continuously. Such re-identification systems, handling large number of previously unseen people, may incorrectly assign separate labels to different instances of the same person. Instead of relying on the user set threshold, we propose an optimization framework with the possibility of encountering previously unseen persons.

With a continuous inflow of data, a human in the loop is queried for labels as well as discriminating attributes to update itself and build a knowledge base about the attributes. Starting with absolutely no attribute information, the system uses the incrementally built knowledge base to reduce the burden on the human as time progresses. This approach

makes the system capable of selecting useful attributes without being restricted by any predetermined set of attributes to start with. We validate the performance of the proposed method using two publicly available benchmark datasets - WARD [80] and i-LIDS-VID [112] and compare with state-of-the-art re-identification methods.

The rest of the chapter is organized as follows. Section 4.2 briefly discusses the related works. An overview of the proposed approach is given in Section 4.3. The details about the re-identification approach, as active image pair selection, and use of attributes are described in Section 4.4. Experimental results and comparisons are shown in Section 4.5. Finally, conclusions are drawn in Section 4.6.

4.2 Related Works

In the last few years there has been increasing attention in the fields of person re-identification, active and attribute based learning. Since the previous chapters have detailed discussions on some of the related works in person re-identification and active learning, here we will discuss, in brief, about some of the previous works in attribute based learning.

Attributes based learning: The notion of attributes comes from the literature on concepts and categories [84]. While features like HOG, SIFT, LBP *etc.* have dominated in computer vision tasks such as scene classification, object detection *etc.*, these are low level descriptions of objects which is understandable to machine only. On the other hand, attributes are describable aspects of information such as a facial expression, age, gender, pose or could be any other side information such as *'has bagpack', 'has hat'* or *'is the animal furry' etc.* Due to the descriptive nature of the attributes, these work as excellent communication tools between human and machines to facilitate boosted learning experience by using attributes to provide feedback to different models [61, 89]. Some recent person re-

identification approaches used attributes mainly as a replacement of low level features [64, 65, 73, 105]. While these works use pre-annotated data with a predefined vocabulary of such attributes, our proposed work uses a set of useful attributes with active labeling from the human in the loop.

4.3 Overview of Proposed Approach

The person re-identification system is based on a low level feature based multi-class classifier where each class corresponds to a separate person. To get started, the classifier is trained on a small amount of training data, labeled only with the person ids without any attribute information.

As the next batch of unlabeled images arrives, the feature based classifier chooses the most informative unlabeled image (query image) and a list of candidate images (sample images) from the labeled set. The query-sample pair is chosen so that subsequent misclassification risk is maximally reduced upon getting the label of the query image from the human expert. The human expert labels the query image by answering 'yes' or 'no' to the question whether the query and sample image are of the same person or not. The probable sample images for a particular query image are presented to the expert in decreasing order of the sample image's class membership probabilities. The class membership probability distribution, given by the classifier, expresses the probability of the query image to belong to one of the already labeled persons from where the sample images are chosen. Section 4.4.1 discusses in detail the underlying principle behind choosing the active image pair selection inducing maximal information upon labeling the query image.

Along with the match-mismatch response, the expert gives the most appropriate attribute level explanation for each mismatch (*e.g.*, these two images don't match as one

has blue shirt while the other does not). While the expert given label is used to update the person re-identification system, the attribute feedback is used to create a knowledge base to be employed in two advantages. Firstly, the knowledge base helps to learn a set of attribute predictors which can, subsequently, be used to get better estimate of the class membership probabilities. Thus, the system can start without predetermined attribute vocabulary and pretrained attribute predictors requiring tedious labeling. Secondly, the attribute knowledge base helps in reducing the search space by removing improbable sample images having similar attribute characteristic. This reduces the burden on the human expert as he/she has to compare less number images to obtain the label of the query. Section 4.4.2 discusses in detail the use of attribute feedback for continuous person re-identification.

4.4 Methodology

In this section, our proposed solution towards continuous learning of person re-identification is discussed in details.

4.4.1 Active Image Pair Selection

As described in the previous section, the query is selected such that the expert's response is likely to reduce the subsequent misclassification risk the most at the cost of least annotation cost. Let the number of labeled classes at a certain moment be K and the K length class membership distribution of an unlabeled image x be $\mathbf{p}_x = \{p_x^1, p_x^2, \cdots, p_x^K\}$. For the sake of simplicity let us assume equal risk if x is misclassified into any of the classes other than its true class. x can belong to a previously seen or an unseen class. Let us consider the case when x is an image from a person seen previously. The estimated misclassification risk for such an example is given by,

$$\mathbf{R}_{oc}(x) = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} p_x^i \cdot p_x^j \tag{4.1}$$

where, the subscript '*oc*' in $\mathbf{R}_{oc}(x)$ denotes that x is an image from an old class already belonging to the set of labeled classes. In a similar fashion, when x belongs to a previously unseen class, the estimated misclassification risk is,

$$\mathbf{R}_{nc}(x) = \sum_{j=1}^{K} p_x^j = 1 \tag{4.2}$$

where, the subscript '*nc*' denotes new class. Let $P_n(x)$ be the probability that x is a previously unseen person. In that case, the total expected misclassification risk is given by,

$$\begin{aligned} \mathbf{R}(x) &= \big(1 - P_n(x)\big)\mathbf{R}_{oc}(x) + P_n(x)\mathbf{R}_{nc}(x) \\ &= \mathbf{R}_{oc}(x) + P_n(x)\big(1 - \mathbf{R}_{oc}(x)\big) \end{aligned} \tag{4.3}$$

Ideally, the class membership distribution of an image of a previously unseen person will be more uncertain than an image of a person seen previously. *Shannon entropy* is a measure of uncertainty of an event characterized by its probability distribution. For any image x with class membership distribution $\mathbf{p}_x$, the entropy is given by $H(x) = -\sum_{i=1}^{K} p_x^i \ln p_x^i$. The probability of being a new class can be estimated as a fraction of its entropy compared to the maximum entropy which occurs when the class membership distribution is most uncertain. The maximum value of entropy of an event is characterized by an uniform distribution and the entropy is given by $\ln K$. Thus, $P_n(x)$ is given by,

$$P_n(x) = \frac{-\sum_{i=1}^{K} p_x^i \ln p_x^i}{\ln K} = \frac{\sum_{i=1}^{K} p_x^i \ln \frac{1}{p_x^i}}{\ln K} \tag{4.4}$$

Using this value of $P_n(x), \mathbf{R}_{oc}(x)$ and $\mathbf{R}_{nc}(x)$ from equations (4.4), (4.1) and (4.2) in eqn. (4.3), the misclassification risk of x can be expressed in terms of the class membership

probabilities as,

$$\mathbf{R}(x) = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} p_x^i \cdot p_x^j + \frac{\sum_{i=1}^{K} p_x^i \ln \frac{1}{p_x^i}}{\ln K} \Big(1 - \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} p_x^i \cdot p_x^j\Big) \tag{4.5}$$

Once a query is obtained, the samples from the labeled set of images are presented to the human according to their chances of match to the query. To avoid notational complexity, let us assume that the class membership distribution $\{p_x^1, p_x^2, \cdots, p_x^K\}$ is sorted in order of decreasing value. Sample image from class 1 will be presented to the expert first, then, the sample image from class 2 and so on. Thus, p_x^i also gives the probability of getting a match for x in *exactly* i comparisons. As the expert has to either accept or deny the chosen sample image, the cost of labeling the query, essentially, is proportional to the number of sample images presented before a match is found. Since p_x^i denotes the probability of getting a match in exactly i comparisons, the expected number of comparisons $\mathbf{C}(x)$ is given by,

$$\mathbf{C}(x) = \sum_{i=1}^{K} p_x^i \cdot i \tag{4.6}$$

The optimum query x^* is to be selected such that on labeling x^*, misclassification risk is reduced maximally at the cost of minimum number of comparisons. Mathematically,

$$x^* = \underset{x}{\operatorname{argmax}} \big(\mathbf{R}(x) - \mathbf{C}(x)\big) \tag{4.7}$$

Fig. 4.1 summarizes the proposed active image pair selection framework for continuous person re-identification.

4.4.2 Use of Attribute Feedback

The role of the human in the loop is further extended in the sense that our model learns the way human uses different traits or attributes (*e.g.*, 'having long hair' or 'wearing green colored shirt or not') to discriminate between persons. The attribute information

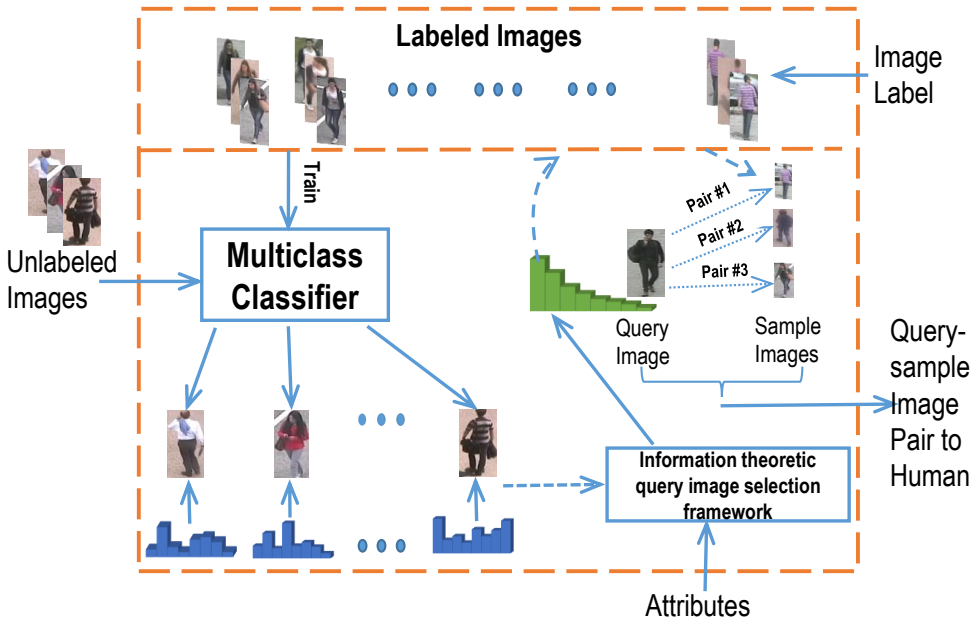


Figure 4.1: Image pair selection. Samples are presented in order of decreasing probabilities obtained from the sorted class membership distribution of the query image. Next the query-sample pair is examined by the expert for match.

about the already labeled persons is used to choose or discard sample images on top of the order determined by the class membership probabilities. Fig. 4.2 shows the high level scheme of such use of attributes with active learning. To keep the burden on the human expert to minimum, only the attributes which distinctly discriminates the query and the sample are sought. For a match, finding attributes which differentiates the person from all others is harder. As a result, the human expert is asked to give attribute feedback only for non-matches.

Assume, for a mismatch, the expert identifies the attribute a_q as not present in the query image x while it is present in the sample image from class k. This information is stored against the respective classes in an attribute knowledge base. A short term advantage of the knowledge base is that, before choosing the next image, classes having the same trait as class k with respect to the attribute a_q are removed from being a match to x. This

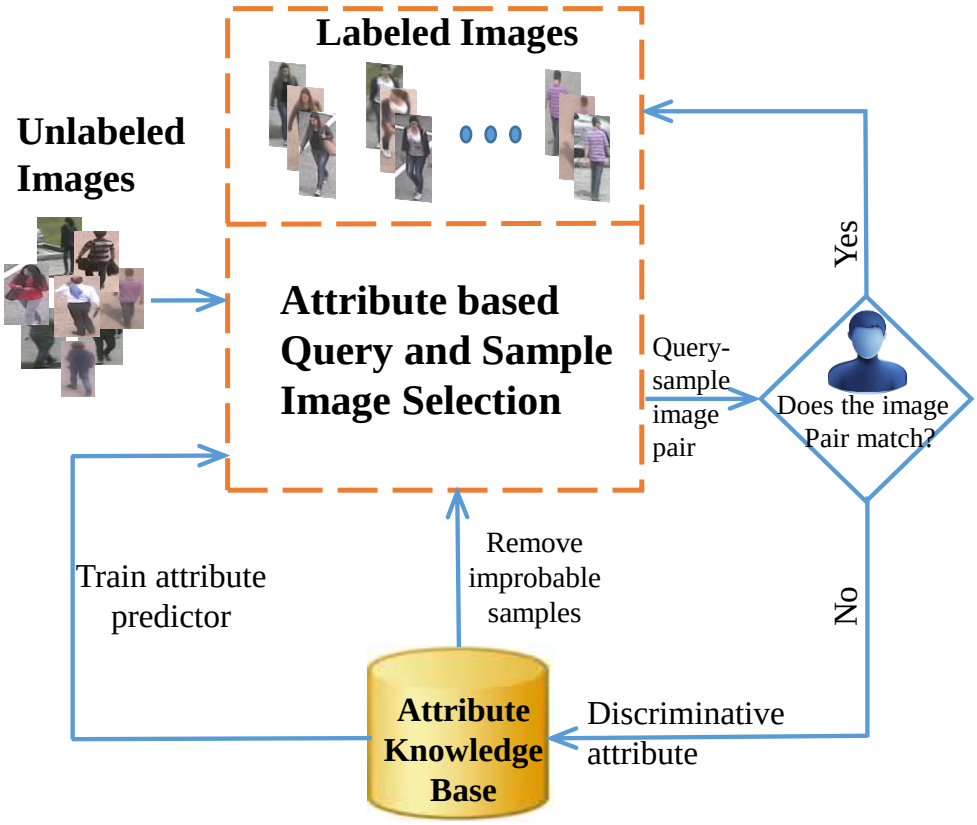


Figure 4.2: Attribute feedback in image pair selection. As unlabeled images come, the classifier along with an attribute predictor, learned on the way, selects a query image which is presented to the human along with candidate matches from the labeled pool. The human does the labeling and gives attribute based explanation of the mismatches, that, in turn, is used to learn and improve the attribute predictors.

reduces the annotation cost as the expert does not have to judge repetitively on sample images with similar attributes.

Another advantage of the attribute feedback is that it helps in reducing the number of comparisons by building attribute predictors on the way based on this acquired knowledge. Let a set of M binary attribute predictors are trained on M different attributes. Each of the predictors gives a $\{1, 0\}$ output where 1 implies the presence of the attribute and 0 implies otherwise. Let $\mathbf{A}_k = \{a_1^k, a_2^k, \cdots a_M^k\}$ be the set containing the attribute labels for images of class k with an index set $\mathbf{I}_k \subset \{1, 2, \cdots M\}$. $\mathbf{I}_k$ contains the attribute indices which got labels from the expert at any moment for this class. Elements of the set $\mathbf{A}_k$ is defined as,

$$a_i^k \in \begin{cases} \{\phi\} & \text{if } i \notin \mathbf{I}_k, \quad [\{\phi\} \text{ denotes a null set}] \\ \{0,1\} & \text{otherwise} \end{cases} \tag{4.8}$$

Let the number of such labeled attributes be m_k (*i.e.*, $|\mathbf{A}_k| = m_k$). Let the class membership probabilities of the unlabeled image x provided by the re-identification system be denoted as $\{p_{x,r}^1, p_{x,r}^2, \cdots p_{x,r}^K\}$. These probabilities are modified by running the attribute predictors on x for the attributes in $\mathbf{A}_k$. Let m_x of the predicted attribute values of x match with the corresponding attributes of class k. We employ a majority voting strategy to refine these class membership probabilities to get p_x^k as,

$$p_x^k = \begin{cases} p_{x,r}^k \cdot e^{\frac{m_x}{m_k}} & \text{if } m_x > m_k - m_x \\ p_{x,r}^k \cdot e^{-\frac{m_k - m_x}{m_k}} & \text{if } m_x < m_k - m_x \\ p_{x,r}^k & \text{otherwise} \end{cases} \tag{4.9}$$

The refined class membership probability values are used (Section 4.4.1) to select the most informative query for labeling.

4.5 Experiments

To validate our approach, we performed experiments on two benchmark datasets - WARD [80] and iLIDS-VID [112]. Some of the popular datasets (*e.g.*, VIPER), though, have more persons, the number of images per person is too few to suit a continuous framework.

Objective: The main objective of the experiments is to analyze how well the proposed framework is capable of updating itself with continuous inflow of data in terms of new as well as old persons. In an active learning set up, we want to see how the feedback about the attributes from the human expert helps in reducing annotation effort. Towards

these goals, we compare the performance of our framework with the following two baselines. **Baseline-I** assumes that no attribute information is fed back by the expert while comparing unlabeled images with labeled samples. **Baseline-II** assumes that information about every attribute of every unlabeled image is provided as feedbacks. These two baselines are two extremes where the former assumes no attribute information and the later assumes perfect attribute information for all labeled images. The proposed framework, on the other hand, uses attribute predictors which is incrementally built based on the attribute feedback. This scenario lies in between the two baselines and is validated by the experimental results. The attributes used for baseline II are listed in appendix A.

Experimental Setup:

- Images have been normalized to 128×64 to be consistent with the state-of-the-art person re-identification methods.
- Mean color feature (HSV) is generated following the scheme in [45]. Before computing these features, three salient regions (head, torso and legs) are extracted from the images as described in [7]. The head region is discarded, since it often consists of a few and less informative pixels. We additionally divide both body and torso into two horizontal sub-regions based on the intuition that people can wear shorts or long pants, and short or long sleeves tops. Each bodypart, is divided into blocks of size 8×16 and the blocks are overlapping by 50% in horizontal and vertical directions.
- Both the dataset are divided into 4 batches so that 25 % of the total persons are seen for the first time. Along with the new persons, each batch also contains images from 50% of the persons seen till the previous batch. As a concrete example, the first batch of the WARD dataset contains images for 18 people (approximately 25% of the total 70 persons). The second batch contains images of new 18 persons as well as 9 old persons. The third batch, similarly, contains images of new 18 persons as well as 18

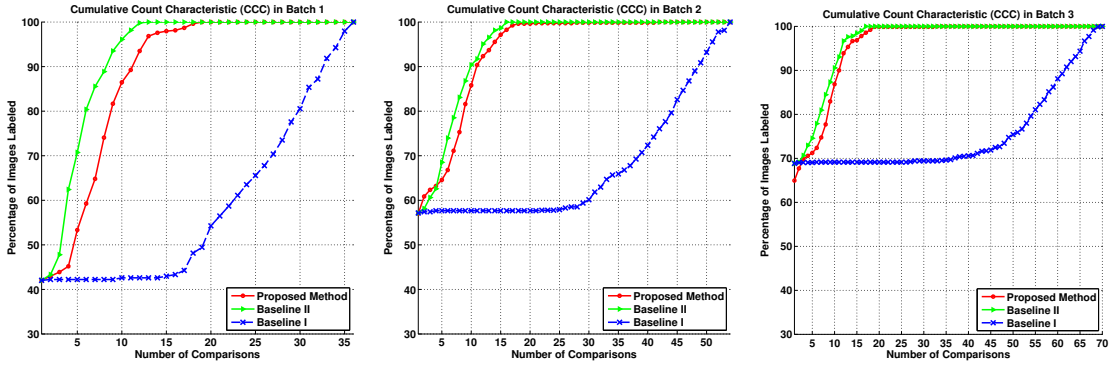
old persons and so on. The initial training is done on the first batch assuming labeled data but no attribute information. The disjoint test set is created using 2 images per person per camera. We ran 5 independent trials for each test and report the average results.

- We use a linear Support Vector Machine (SVM) throughout, as the multi-class classifier for person re-identification and the binary classifier for attribute prediction. The toolbox LIBSVM [16] is used for the experimentations.

Performance Measures: Results are evaluated in terms of the number of comparisons to get labels for all the images in each batch. This is shown as a Cumulative Count Curve (CCC) which gives the number of images (%) getting labeled within a certain number of comparisons. As an example, say the number of unlabeled images getting labeled after exactly the first and second comparison be 10 and 5 respectively. So cumulatively the number of images getting labeled within a maximum of 2 tries is 10+5 = 15. The CCC plot, in that case, has 1 and 2 in the x axis corresponding to 10 and 15 in the y axis. As the number of classes vary in each batch we express the y axis in percentage. Another metric that is compared is the number of persons labeled vs the number of binary comparisons. For each batch, we also provide the total number of comparisons and the comparisons per image to get all the images labeled. We also provide the accuracy of the person re-identification system as more and more batches of data get labeled. All the comparisons were either taken from the published results or by running codes which are publicly available.

4.5.1 WARD Dataset

The WARD dataset [80], as also used in the previous chapter, has 4786 images of 70 different people acquired in a real surveillance scenario in three non-overlapping cameras. The initial training was done using 18 of the 70 persons while the 3 subsequent batches had

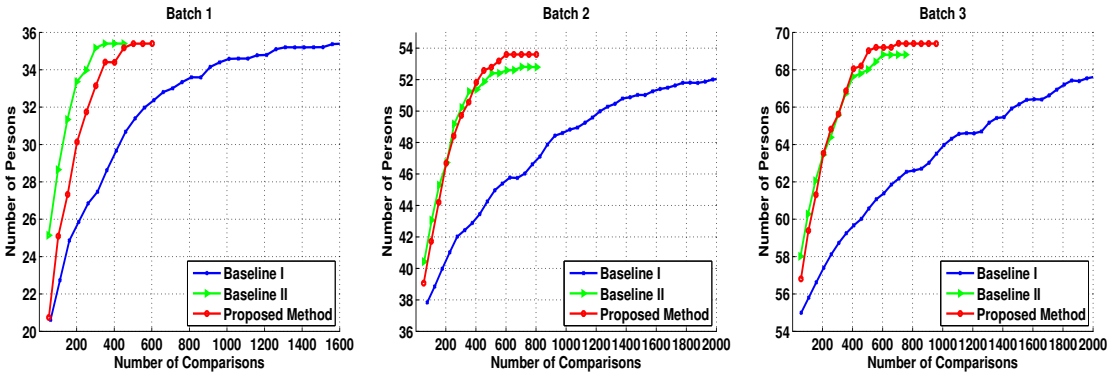


(a) (b) (c)

Figure 4.3: CCC curves for the WARD dataset. Comparison count performances in (a), (b) and (c) are shown for batch 1, 2 and 3 respectively.

18, 18 and 16 previously unseen persons respectively. These 3 batches also contained 9, 18 and 27 previously seen persons (randomly chosen). Fig. 4.3(a), (b) and (c) compare the percentage cumulative count of the proposed method with the 2 baselines for batch 1, 2 and 3 respectively.

It can be seen that as more and more data comes, more and more images are labeled with smaller number of comparisons by the expert. In both the baselines 42.03%, 57.16% and 68.85% of the unlabeled images are presented with the true class image as the very first sample image in batch 1, 2 and 3 respectively. That is, 42.03%, 57.16% and 68.85% unlabeled images get their labels within the first comparison. For the rest of the images the number of comparisons increase gradually for baseline I (using no attribute information) such that it takes upto 36, 54 and 68 comparisons per image to get 99% of the images labeled. These numbers are 12, 16 and 16 when all attribute information are known (baseline II) while the same numbers for the proposed method are 18, 17 and 18 for batch 1, 2 and 3 respectively. It should be noted that the slightly better performance of baseline II comes at the cost of much more effort from the human expert as this requires all the attributes to be labeled by him/her for each person. In batch 3, the number of



(a) (b) (c)

Figure 4.4: Comparison of the labeling effort of the proposed method with the two baselines in terms of the number of persons labeled vs the number of comparisons to get these many persons labeled. (a), (b) and (c) shows the comparative performance for batch 1, 2 and 3 respectively for the WARD dataset. For convenience of visualization, the plot for Baseline I is not shown till the end.

images getting labeled within the first comparison for the proposed method is little less than both the baselines (64.95% vs 68.85%). This is due to the fact that, the uncertainty in the attribute predictor affects the class membership distribution of some of the unlabeled images badly such that the probability of true class decreases. But, the catching up of the proposed method with baseline II suggests that the attribute information helps to get more number of images labeled with little effort while affecting a few by increasing the number of comparisons.

Fig. 4.4(a), (b) and (c) show comparative plots of number of persons being labeled vs the number of binary comparisons for batch 1, 2 and 3 respectively. With increasing batch number the proposed method performs close to the Baseline II whereas Baseline I takes much more number of comparisons to get the same number of persons labeled.

Table 4.1 gives a comparative analysis of the total and average number (per image) of comparisons to label all the unlabeled images for each batch. Similar to WARD, the proposed method performs the best in terms of the average number of comparisons with increasing images and classes. As the number of images in each batch also increases, the

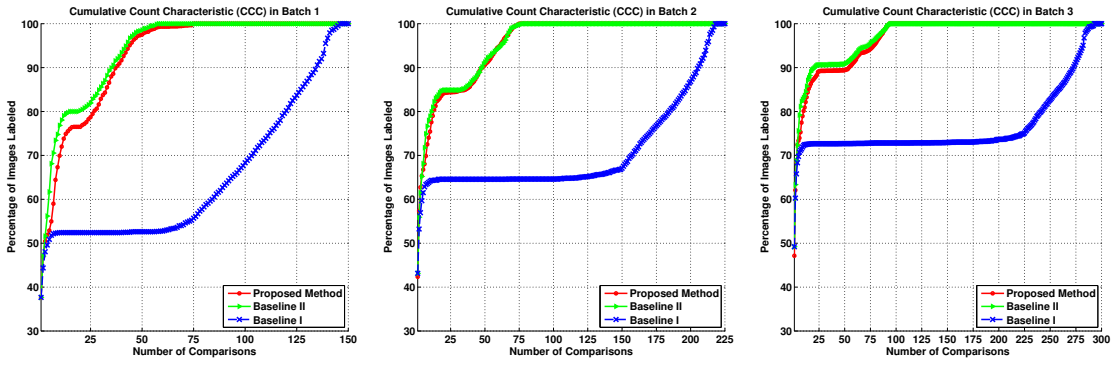
Table 4.1: Total and average number (per image) of query-sample pair comparisons made by the expert to get all the images labeled. For both the datasets the proposed method is close to baseline II. Baseline I requires far more number of comparisons to get all the images labeled than the other two methods. The numbers are larger in case of the i-LIDS-VID as the number of people is more in this dataset than WARD.

		batch 1	batch 2	batch 3
WARD	Baseline I	1720.4	3029.4	3846.6
	Baseline I (avg)	15.9	18.7	18.3
	Proposed	**577.4**	**727.8**	**827.4**
	Proposed (avg)	**5.3**	**4.5**	**3.9**
	Baseline II	422.6	648.2	713.8
	Baseline II (avg)	3.9	4.0	3.4
i	Baseline I	16046.8	29868.8	42437.4
LIDS	Baseline I (avg)	148.6	184.4	202.1
-VID	Proposed	**3517.2**	**5123.6**	**6620**
	Proposed (avg)	**32.6**	**31.6**	**31.5**
	Baseline II	2960.4	4920.2	5797.4
	Baseline II (avg)	27.4	30.4	27.6

average number of comparisons per image to label all of them is also provided. We see that the proposed method reduces the effort of the expert considerably by using attribute information and is close to baseline II where all attributes of the labeled images are known. In terms of average number of comparisons, the proposed method is the best among the 3 as it decreases gradually even if the total number of both images and classes increase from batch 1 to 3.

4.5.2 i-LIDS-VID Dataset

iLIDS-VID [112] is a recently introduced person re-identification dataset. This dataset consists of images from 300 people at an airport arrival hall captured through 2 non-overlapping cameras. Apart from a large number of people, the challenges in this dataset also includes clothing similarities, clutter and lighting variations among others. For

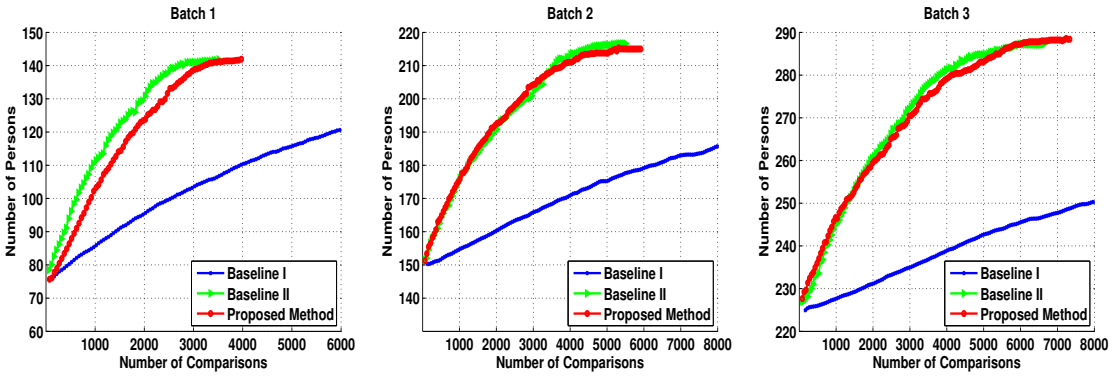


(a) (b) (c)

Figure 4.5: CCC curves for the i-LIDS-VIDS dataset. Comparison count performances in (a), (b) and (b) are shown for batch 1, 2 and 3 respectively.

this dataset, the initial training was done using 75 of the 300 persons while the 3 subsequent batches had 75 previously unseen persons each. These 3 batches also contained respectively 37, 74 and 111 randomly chosen previously seen persons. Fig. 4.5(a), (b) and (c) compare the percentage cumulative count of the proposed method with the 2 baselines for batch 1, 2 and 3 respectively.

Similar to the WARD dataset, i-LIDS-VID comprising of a lot more persons, shows that with more and more data, number of images getting labeled in the very first comparison increases. For both the baselines 37.65%, 43.14% and 49.19% of the unlabeled images are presented with the true class image as the very first sample image in batch 1, 2 and 3 respectively. That is, 37.65%, 43.14% and 49.19% unlabeled images get their labels within the first comparison. The numbers are lower than WARD as the number of persons are much more in this dataset than WARD. For the rest of the images the number of comparisons increase gradually for baseline I such that it takes upto 144, 217 and 288 comparisons per image to get 99% of the images labeled. These numbers are 53, 70 and 90 when all attribute information are known (baseline II) while the same numbers for the proposed method are 57, 71 and 90 for batch 1, 2 and 3 respectively.



(a) (b) (c)

Figure 4.6: Comparison of the labeling effort of the proposed method with the two baselines in terms of the number of persons labeled vs the number of comparisons to get these many persons labeled. (a), (b) and (c) shows the comparative performance for batch 1, 2 and 3 respectively for the i-LIDS-VID dataset. For convenience of visualization, the plot for Baseline I is not shown till the end.

Fig. 4.6(a), (b) and (c) show comparative plots of number of persons being labeled vs the number of binary comparisons for batch 1, 2 and 3 respectively. Similar to WARD, with increasing batch number the proposed method performs close to the Baseline II whereas Baseline I takes much more number of comparisons to get the same number of persons labeled.

Fig. 4.7 and 4.8 provide examples from i-LIDS-VID dataset showing that the proposed method reduces the labeling effort considerably than Baseline I, where no attribute information is used. The sample images are shown from left to right in the order they are presented to the expert according to the decreasing class membership probability value. Fig. 4.7 shows that the proposed method's performance is close to that of Baseline II where all attributes for all the labeled images are available. While baseline I takes 148 (the first 10 are shown only) and baseline II takes 6 sample images, the proposed method is close to Baseline II taking 7 sample images to label the query image. The attribute 'hasbackpack' helps to reduce the labeling effort here. Similarly fig. 4.8 shows one case where the proposed method beats the Baseline II, even though attributes are not labeled explicitly by the expert

Figure 4.7: While Baseline I takes 148 (the first 10 are shown only) and Baseline II takes 6 sample images, the proposed method is close to Baseline II taking 7 sample images to label the query image. The attribute 'hasbackpack' helps to reduce the labeling effort here.

in the proposed method. While, for this case, baseline I takes 142 (the first 10 are shown only) and baseline II takes 9 sample images, the proposed method takes 7 sample images to label the query image. The attribute 'hashandbagcarrierbag' helps to reduce the labeling effort here.

Table 4.2: Comparison of the proposed method with the state-of-the-art in terms of re-identification accuracy (%).

	Batch 1	Batch 2	Batch 3
Proposed	15.87	24.8	31.07
MS-Color&LBP+DVR [112]	-	-	34.5
MS-Color+DVR [112]	-	-	32.7
MS-SDALF [7]	-	-	6.3
MS-SDALF+DVR [112]	-	-	26.7

Figure 4.8: While Baseline I takes 142 (the first 10 are shown only) and Baseline II takes 9 sample images, the proposed method takes 7 sample images to label the query image. The attribute 'hashandbagcarrierbag' helps to reduce the labeling effort here.

Table 4.1 gives a comparative analysis of the total and average number (per image) of comparisons to label all the unlabeled images for each batch. Though, traditional re-identification methods have published re-identification accuracy based on batch training, we report the accuracy on the continuous setting. Since, WARD is a 3 camera dataset and the published results on it are camera pairwise, we can not compare the results on WARD. Though re-identification accuracy of the proposed approach have been reported after each batch of data have been labeled, the comparison with the state-of-the-art can only be done after the framework sees all the persons. Table 4.2 gives such a comparative analysis of the test accuracy. To compare fairly, we test on the persons which the framework has seen till this point. It can be seen that the test accuracy increases gradually to reach the state-of-the-art. It should be noted that the proposed method sees only a few images compared

to [112] and other mulltishot approaches where sequence of images for each person were used for training.

4.6 Conclusions

In this work, we addressed the problem of continuously re-identifying persons starting with a small set of labeled data in an active learning set up. We also showed that mid level attribute based explanations from the expert help in reducing the effort of getting labels for unlabeled images. A set of attribute predictors are also learned online which helps to transfer the domain knowledge of the expert to the model. Experiments on two publicly available benchmark datasets are performed to validate our proposed approach. The future directions of our research will be to apply the framework to bigger networks with large numbers of cameras, and cope with wider space-time horizons in a continuous setting.

Chapter 5

Conclusions

5.1 Summary of the Research Contributions

Networks of vision sensors seem to be the next paradigm for addressing security needs to disaster response to environmental monitoring. Consistent increase of sensor quality at continuously diminishing cost facilitates the coverage of wide area having hundreds of cameras resulting in tens of thousands of hours of videos. Analyzing such a massive volume of data to re-identify persons coming in and out of the non-overlapping field of views in such wide area camera network is challenging. Hence studying transformation of features between cameras and involving human efficiently towards large-scale camera network video analysis are extremely impactful to systems that are starting to be deployed and are gaining importance among the research community. In this dissertation, we have presented a novel framework for studying feature transformation towards robust person re-identification. We have also presented some strategies to reduce human effort in dealing with big data in a continuous re-identification scenario.

Chapter 2 presented a mechanism to re-identification by modeling the way feature gets transformed between cameras. The similarity between the feature histograms and time

series data motivated us to apply the principle of Dynamic Time Warping to model the transformation of features by warping the feature space. After capturing the feature warps, the variabilities of the warp functions were modeled as a function space of feature warps. The function space not only allowed us to model feasible transformation between pairs of instances of the same target, but also to separate them from the infeasible transformations between instances of different targets. We show that our approach is robust with respect to severe illumination and pose variations by evaluating the performance on five datasets.

Chapter 3 explored the option of involving human efficiently in boosting the re-identification performance in an active learning set up with two different goals - reducing the labeling effort and the training time by updating the model online. For large multi-sensor data as typically encountered in person re-identification, labeling lot of samples is not only an overhead but does not always mean more information, due to redundant labeling. We propose a convex optimization based iterative framework that progressively and judiciously chooses a sparse but informative set of samples for labeling, with minimal overlap with previously labeled images. The framework not only helps in reducing the labeling effort but also updates the model online by using a sparse representation based classifier when new unlabeled data arrives continuously. Experiments on three publicly available benchmark datasets are performed to validate the proposed approach.

Chapter 4 extended the idea of involving human efficiently by incorporating domain knowledge from the experts to the active learning mechanism to help the mechanism in helping people by asking informative questions. Mid level attribute based explanations from the human annotator and 'value of information' based binary comparison strategy were used towards this objective. The binary yes-no type questions helped in reducing the effort of the human by avoiding comparison with many categories (persons). We demonstrate the

effectiveness of the proposed method for continuous person re-identification system with two datasets.

5.2 Future Research Directions

5.2.1 Person Re-Identification in Egocentric Videos

With wearable cameras such as the Go-Pro becoming popular, recognizing persons in videos captured from first-person cameras is essential for many applications such as behavior understanding, retrieval, assistive vision technologies *etc.* Apart from the expanding application areas, vision researchers are actively taking part in exploring these so called egocentric videos due to certain advantages such as presence of the object of interest in the center and in focus, less occlusion among others. Recent works use supervised learning to recognize activities [33, 35] social interactions [34] *etc.* Unsupervised methods include scene discovery [49], key frame selection [27] or video summarization [67, 75]. Tracking and re-identification of persons in egocentric videos is a challenging future direction of research.

Different aspects of person re-identification depending on specific application areas makes it a challenging problem in egocentric videos. For multiple wearable cameras mounted in the bodies of multiple agents, consensus among them is necessary when re-identifying or tracking a person. Again for a single wearable system (*e.g.*, analysis of first-person sports videos [54]), it is necessary to re-identify players leaving the camera FOV and re-entering in the same camera but possibly in different viewing angle, scale or resolution. Another application area can be tracking the gaze of a first person viewer to identify salient regions/features on the subject to which the first person viewer focuses while trying to re-identify a person. Embedding this can aid an active learning system by identifying salient regions to look at for re-identification. Again, trained human operators mostly

perform person re-identification by focusing on particular small parts of a person of interest. Helping human annotators in active learning based system, by mimicking the approaches of such trained human operators would be realistically achievable using pan-tilt-zoom control wearable cameras trained to provide selective focus on body parts from a distance. Using such a technology in the active re-identification system can be exploited as future work.

5.2.2 Active Selection with Scene Context

Recent successes in visual recognition take advantage of the fact that, in nature, objects and events tend to co-exist with each other in a particular configuration. This is often termed as *context* and plays an important role in human visual system [86]. Several research works [18, 122] considered the use of context from different perspectives to recognize human activities and showed significant performance improvement over the approaches that do not use context. The question we want to ask here is whether such contextual information could be utilized for choosing the set of examples to be labeled. Most approaches to selecting the examples to be labeled exploit informativeness, expected error reduction (EER), etc. of *individual* data instances in a batch or in an online manner assuming that there are no inter-relationships among them [52, 111]. Few works, such as [4] utilizes the inter-relationship of the data instances in feature space for active learning. However, objects and activities by persons in video exhibit a much richer set of spatial and temporal interactions between themselves and a natural question to explore is whether the contextual relationships between objects, activities and persons aid in selecting subsets of exemplars with high degree of similarity so that labeling a few of them improves the system with maximum effect.

Bibliography

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] Ejaz Ahmed, Michael Jones, and Tim K Marks. An Improved Deep Learning Architecture for Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[3] Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. Reference-Based Person Re-Identification. In *Advanced Video and Signal-Based Surveillance*, 2013.

[4] Oisin Mac Aodha, Neill D F Campbell, Jan Kautz, and Gabriel J Brostow. Hierarchical Subquery Evaluation for Active Learning on a Graph. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[5] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning Implicit Transfer for Person Re-identification. In *European Conference on Computer Vision, Workshops and Demonstrations*, volume 7583 of *Lecture Notes in Computer Science*, pages 381–390, 2012.

[6] Slawomir Bak, Etienne Corvée, Francois Brémond, and Monique Thonnat. Boosted Human Re-identification using Riemannian Manifolds. *Image and Vision Computing*, 30(6-7):443–452, June 2012.

[7] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-Driven Accumulation of Local Features for Human Characterization and Re-identification. *Computer Vision and Image Understanding*, 117(2), 2012.

[8] Loris Bazzani, Marco Cristani, Alessandro Perina, Michela Farenzena, and Vittorio Murino. Multiple-Shot Person Re-identification by HPE Signature. In *International Conference on Pattern Recognition*, pages 1413–1416, August 2010.

[9] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[10] Aurelien Bellet, Amaury Habrard, and Marc Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data. *ArXiv e-prints*, 2013.

[11] Donald J Bemdt and J Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In *Working Notes of the Knowledge Discovery in Databases Workshop*, pages 359–370, 1994.

[12] Claude Berge. *Hypergraphs: Combinatorics of Finite Sets*, volume 45. Elsevier, 1984.

[13] Arijit Biswas and Devi Parikh. Simultaneous Active Learning of Classifiers & Attributes via Relative Feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[14] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[15] Shayok Chakraborty, Vineeth N Balasubramanian, and Sethuraman Panchanathan. Optimal Batch Selection for Active Learning in Multi-label Classification. In *ACM International Conference on Multimedia*, pages 1413–1416, 2011.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[17] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conference*, pages 68.1–68.11, 2011.

[18] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning Context for Collective Activity Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[19] Yang Cong, Junsong Yuan, and Jiebo Luo. Towards Scalable Summarization of Consumer Videos via Sparse Dictionary Selection. *IEEE Transactions on Multimedia*, 14(1):66–75, 2012.

[20] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[21] Abir Das, Anirban Chakraborty, and Amit K. Roy-Chowdhury. Consistent Re-identification in a Camera Network. In *European Conference on Computer Vision*, pages 330–345. Springer, 2014.

[22] Ankur Datta, Lisa M Brown, Rogerio Feris, and Sharathchandra Pankanti. Appearance Modeling for Person Re-Identification using Weighted Brightness Transfer Functions. In *International Conference on Pattern Recognition*, 2012.

[23] Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes, Antonio da Luz, and Arnaldo de Albuquerque Araújo. VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[24] Weihong Deng, Jiani Hu, and Jun Guo. Extended SRC: Undersampled Face Recognition via Intraclass Variant Dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012.

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[26] Mert Dikmen, Emre Akbas, Thomas S Huang, and Narendra Ahuja. Pedestrian Recognition with a Learned Metric. In *Asian Conference on Computer Vision*, pages 501–512, 2010.

[27] Aiden R Doherty, Daragh Byrne, Alan F Smeaton, Gareth J F Jones, and Mark Hughes. Investigating Keyframe Selection Methods in the Novel Domain of Passively Captured Visual Lifelogs. In *International Conference on Content-based Image and Video Retrieval*, pages 259–268. ACM, 2008.

[28] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. In *Advances in Neural Information Processing Systems*, pages 19–27, 2012.

[29] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607, 2012.

[30] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and Shankar S Sasrty. A Convex Optimization Framework for Active Learning. In *IEEE International Conference on Computer Vision*, pages 209–216, 2013.

[31] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and Appearance for Mobile Scene Analysis. In *IEEE International Conference on Computer Vision*, pages 1–8, October 2007.

[32] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing Objects by their Attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[33] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding Egocentric Activities. In *IEEE International Conference on Computer Vision*, pages 407–414, 2011.

[34] Alireza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A Firstperson Perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–327, 2012.

[35] Alireza Fathi, Yin Li, and James M Rehg. Learning to Recognize Daily Action using Gaze. In *European Conference on Computer Vision*, pages 314–327, 2012.

[36] Hans G Feichtinger and Thomas Strohmer. *Gabor Analysis and Algorithms: Theory and Applications*. Springer Publications, 1998.

[37] Brendan J Frey and Delbert Dueck. Clustering by Passing Messages between Data Points. *Science*, 315(5814):972–976, 2007.

[38] Shenghua Gao, Ivor Wai-Hung Tsang, and Liang-Tien Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92–104, 2013.

[39] Salvador García, J Derrac, J R Cano, and F Herrera. Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3), 2012.

[40] Andrew Gilbert and Richard Bowden. Tracking Objects across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity. In *European Conference Computer Vision*, pages 125–136, 2006.

[41] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating Appearance Models for Recongnition, Reacquisition and Tracking. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, October 2007.

[42] Douglas Gray and Hai Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *European Conference on Computer Vision*, pages 262–275, 2008.

[43] Marko Heikklä and Matti Pietikäinen. A Texture-based Method for Modeling the Background and Detecting Moving Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):657–62, April 2006.

[44] Martin Hirzer, Peter M Roth, and Horst Bischof. Person Re-identification by Efficient Impostor-Based Metric Learning. In *Advanced Video and Signal-Based Surveillance*, pages 203–208, 2012.

[45] Martin Hirzer, Peter M Roth, Martin Kostinger, and Horst Bischof. Relaxed Pairwise Learned Metric for Person Re-identification. In *European Conference on Computer Vision*, 2012.

[46] H Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[47] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *Advances in Neural Information Processing Systems*, pages 892–900, 2010.

[48] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling Inter-camera Spacetime and Appearance Relationships for Tracking across Non-overlapping Views. *Computer Vision and Image Understanding*, 109(2):146–162, February 2008.

[49] Nebojsa Jojic, Alessandro Perina, and Vittorio Murino. Structural epitome: A way to summarize ones visual experience. In *Advances in Neural Information Processing Systems*, pages 1027–1035, 2010.

[50] Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. Scalable Active Learning for Multiclass Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2259–2273, 2012.

[51] Jean-Claude Junqua and Jean-Paul Haton. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic, 1995.

[52] Vasiliy Karasev, Avinash Ravichandran, and Stefano Soatto. Active Frame , Location , and Detector Selection for Automated and Manual Video Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[53] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact Indexing of Dynamic Time Warping. *Knowledge and Information Systems*, 7(3):358–386, May 2004.

[54] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3248, 2011.

[55] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification. In *British Machine Vision Conference*, 2010.

[56] Krishna Reddy Konda, Andrea Rosani, Nicola Conci, and Francesco G B De Natale. Smart Camera Reconfiguration in Assisted Home Environments for Elderly Care. In *Computer Vision-ECCV 2014 Workshops*, pages 45–58. Springer, 2014.

[57] M Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[58] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[59] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color Invariants for Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[60] L J P Van Der Maaten, E O Postma, and H J Van Den Herik. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, 10(February):1–41, 2009.

[61] Shrenik Lad and Devi Parikh. Interactively Guiding Semi-Supervised Clustering via Attribute-based Explanations. In *European Conference on Computer Vision*, 2014.

[62] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[63] Agata Lapedriza, Hamed Pirsiavash, Zoya Bylinskii, and Antonio Torralba. Are all training examples equally valuable? *arXiv preprint arXiv:1311.6510*, 2013.

[64] Ryan Layne, Timothy Hospedales, and Shaogang Gong. Person Re-identification by Attributes. In *British Machine Vision Conference*, 2012.

[65] Ryan Layne and Timothy M Hospedales. Re-id : Hunting Attributes in the Wild. In *British Machine Vision Conference*, 2014.

[66] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient Sparse Coding Algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2006.

[67] Yong Jae Lee and Kristen Grauman. Predicting Important Objects for Egocentric Video Summarization. *International Journal of Computer Vision*, pages 1–18, 2015.

[68] Thomas Leung and Jitendra Malik. Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[69] Wei Li and Xiaogang Wang. Locally Aligned Feature Transforms across Views. In *International Conference on Computer Vision and Pattern Recognition*, 2013.

[70] Wei Li, Rui Zhao, and Xiaogang Wang. Human Reidentification with Transferred Metric Learning. In *Asian Conference on Computer Vision*, pages 31–44, 2012.

[71] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[72] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person Re-identification : What Features Are Important ? In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 391–401, Florence, Italy, 2012. Springer Berlin Heidelberg.

[73] Xiao Liu, Mingli Song, Qi Zhao, Dacheng Tao, Chun Chen, and Jiajun Bu. Attribute-restricted Latent Topic Model for Person Re-identification. *Pattern Recognition*, 45(12):4204–4213, December 2012.

[74] David G Lowe. Object Recognition from Local Scale-invariant Features. In *IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[75] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013.

[76] Ulrike Von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[77] Bingpeng Ma, Yu Su, and Frederic Jurie. BiCov: a Novel Image Representation for Person Re-identification and Face Verification. *British Machine Vision Conference*, pages 57.1–57.11, 2012.

[78] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local Descriptors Encoded by Fisher Vectors for Person Re-identification. In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 413–422, Florence, Italy, 2012.

[79] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Re-Identification in the Function Space of Feature Warps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1656–1669, 2015.

[80] Niki Martinel and Christian Micheloni. Re-identify People in Wide Area Camera Network. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 31–36, Providence, RI, June 2012.

[81] Alexis Mignon and Frederic Jurie. PCCA : A New Approach for Distance Learning from Sparse Pairwise Constraints. In *International Conference on Computer Vision and Pattern Recognition*, pages 2666–2672, 2012.

[82] Muhammad Mubashir, Ling Shao, and Luke Seed. A Survey on Fall Detection: Principles and Approaches. *Neurocomputing*, 100:144–152, 2013.

[83] Meinard Müller. Dynamic Time Warping. In *Information Retrieval for Music and Motion*, volume 2, chapter 4, pages 69–84. Springer Publications, Berlin, 2007.

[84] Gregory Leo Murphy. *The Big Book of Concepts*. MIT press, 2002.

[85] T Ojala, M Pietikainen, and T Maenpaa. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, July 2002.

[86] Aude Oliva and Antonio Torralba. The Role of Context in Object Recognition. In *Trends in Cognitive Science*, 2007.

[87] Devi Parikh and Kristen Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[88] Devi Parikh and Kristen Grauman. Relative Attributes. In *IEEE International Conference on Computer Vision*, 2011.

[89] Amar Parkash and Devi Parikh. Attributes for Classifier Feedback. In *European Conference on Computer Vision*, 2012.

[90] Elżbieta Pękalska and Robert P W Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*, volume 64. World Scientific, 2005.

[91] Sateesh Pedagadi, James Orwell, and Sergio Velastin. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[92] Robi Polikar, Lalita Udpa, Satish S Udpa, and Vasant Honavar. Learn ++ : An Incremental Learning Algorithm for supervised neural networks. *IEEE Transactions on Systems, Man, and Cybernatics -C*, 31(4):497–508, 2001.

[93] Fatih Porikli and Murray Hill. Inter-Camera Color Calibration Using Cross-Correlation Model Function. In *IEEE International Conference on Image Processing*, pages 133–136, 2003.

[94] Bryan Prosser, Shaogang Gong, and Tao Xiang. Multi-camera Matching using Bi-Directional Cumulative Brightness Transfer Functions. In *British Machine Vision Conference*, September 2008.

[95] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person Re-Identification by Support Vector Ranking. In *British Machine Vision Conference*, pages 21.1–21.11. British Machine Vision Association, 2010.

[96] Qiang Qiu, Zhuolin Jiang, and Rama Chellappa. Sparse Dictionary-based Representation and Recognition of Action Attributes. In *IEEE International Conference on Computer Vision*, pages 707–714, 2011.

[97] S Salvador and P Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. In *KDD Workshop on Mining Temporal and Sequential Data*, 2004.

[98] Riccardo Satta, Giorgio Fumera, and Fabio Roli. A General Method for Appearance-Based People Search Based on Textual Queries. In *European Conference on Computer Vision Workshops*, pages 453–461, 2012.

[99] Riccardo Satta, Giorgio Fumera, and Fabio Roli. Fast Person Re-identification Based on Dissimilarity Representations. *Pattern Recognition Letters*, 33(14):1838–1848, oct 2012.

[100] Cordelia Schmid. Constructing Models for Content-based Image Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–39–II–45, 2001.

[101] William Robson Schwartz and Larry S Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Brazilian Symposium on Computer Graphics and Image Processing*, pages 322–329. IEEE, October 2009.

[102] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[103] Clemens Siebler, Bernardin Keni, and Rainer Stiefelhagen. Adaptive Color Transformation for Person Re-identification in Camera Networks. In *International conference on Distributed Smart Cameras*, number April, pages 199–205, 2010.

[104] Bi Song and Amit K. Roy-Chowdhury. Robust tracking in a camera network: A multi-objective optimization framework. *Selected Topics in Signal Processing*, 2(4):582–596, 2008.

[105] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-Based People Search in Surveillance Environments. In *IEEE Winter Conference on Applications of Computer Vision*, 2009.

[106] Ashok Veeraraghavan, Amit K Roy-Chowdhury, and Rama Chellappa. Matching Shape Sequences in Video with Applications in Human Movement Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1896–1909, December 2005.

[107] Ashok Veeraraghavan, Anuj Srivastava, Amit K Roy-Chowdhury, and Rama Chellappa. Rate-invariant Recognition of Humans and Their Activities. *IEEE Transactions on Image Processing*, 18(6):1326–1339, June 2009.

[108] Alexander Vezhnevets, Joachim M Buhmann, and Vittorio Ferrari. Active Learning for Semantic Segmentation with Expected Change. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3162–3169, 2012.

[109] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People Re-identification in Surveillance and Forensics: a Survey. *ACM Computing Surveys*, 46(2), 2014.

[110] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-Scale Live Active Learning : Training Object Detectors with Crawled Data and Crowds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[111] Carl Vondrick and Deva Ramanan. Video Annotation and Tracking with Active Learning. In *Advances in Neural Information Processing Systems*, 2011.

[112] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-Identification by Video Ranking. In *European Conference on Computer Vision*, 2014.

[113] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[114] Yang Wu, Michihiko Minoh, Masayuki Mukunoki, Wei Li, and Shihong Lao. Collaborative Sparse Approximation for Multiple-Shot Across-Camera Person Re-identification. In *Advanced Video and Signal-Based Surveillance*, pages 209–214. IEEE, September 2012.

[115] Liu Yang and Rong Jin. Distance Metric Learning : A Comprehensive Survey. Technical report, Michigan State University, 2006.

[116] Yang Yang, Jimei Yang, Junjie Yan, Shegcai Liao, Dong Yi, and Stan Z. Li. Salient Color Names for Person Re-identification. In *European Conference on Computer Vision*, pages 536–551, 2014.

[117] Guanwen Zhang, Yu Wang, Jien Kato, Takafumi Marutani, and Mase Kenji. Local Distance Comparison for Multiple-shot People Re-identification. In *Asian conference on Computer Vision*, volume 7726 of *Lecture Notes in Computer Science*, pages 677–690, 2013.

[118] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised Salience Learning for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[119] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph Regularized Sparse Coding for Image Representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[120] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Re-identification by Relative Distance Comparison. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, June 2013.

[121] Xiangxin Zhu, Carl Vondrick, Deva Ramanan, and Charless Fowlkes. Do We Need More Training Data or Better Models for Object Detection? In *British Machine Vision Conference*, volume 3, page 5, 2012.

[122] Yingying Zhu, Nandita M Nayak, and Amit K Roy-chowdhury. Context-Aware Modeling and Recognition of Activities in Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Appendix A

List of Attributes Used for Baseline II

A.1 Atributes information for WARD dataset

Table A.1: List of attributes used for WARD dataset. Information about the bodyparts from where the features are extracted to train the respective attribute predictors are also provided

Attribute	**Features Extracted from**	**Attribute**	**Features Extracted from**
redshirt	upperTorso lowerTorso	patterned	upperTorso lowerTorso
blueshirt	upperTorso lowerTorso	darkbottoms	upperLegs lowerLegs
greenshirt	upperTorso lowerTorso	shorts	upperLegs lowerLegs
darkshirt	upperTorso lowerTorso	hasbackpack	upperTorso lowerTorso
fuchsiashirt	upperTorso lowerTorso	hashandbag- carrierbag	upperLegs

A.2 Atributes information for i-LIDS-VID dataset

Table A.2: List of attributes used for i-LIDS-VID dataset. Information about the bodyparts from where the features are extracted to train the respective attribute predictors are also provided

Attribute	**Features Extracted from**	**Attribute**	**Features Extracted from**	**Attribute**	**Features Extracted from**
redshirt	upperTorso lowerTorso	patterned	upperTorso lowerTorso	yellowshirt	upperTorso lowerTorso
blueshirt	upperTorso lowerTorso	darkbottoms	upperLegs lowerLegs	aquashirt	upperTorso lowerTorso
greenshirt	upperTorso lowerTorso	shorts	upperLegs lowerLegs	skirt	upperLegs lowerLegs
darkshirt	upperTorso lowerTorso	hasbackpack	upperTorso lowerTorso		
fuchsiashirt	upperTorso lowerTorso	hashandbag-carrierbag	upperLegs		