Minimising Human Annotation for Scalable Person Re-Identification

Hanxiao Wang

Submitted in partial fulfilment of the requirement for the Doctor of Philosophy

School of Electronic Engineering and Computer Science

Queen Mary University of London

27 September 2017

Minimising Human Annotation for Scalable Person Re-Identification

Hanxiao Wang

Abstract

Among the diverse tasks performed by an intelligent distributed multi-camera surveillance system, person re-identification (re-id) is one of the most essential. Re-id refers to associating an individual or a group of people across non-overlapping cameras at different times and locations, and forms the foundation of a variety of applications ranging from security and forensic search to quotidian retail and health care. Though attracted rapidly increasing academic interests over the past decade, it still remains a non-trivial and unsolved problem for launching a practical re-id system in real-world environments, due to the ambiguous and noisy feature of surveillance data and the potentially dramatic visual appearance changes caused by uncontrolled variations in human poses and divergent viewing conditions across distributed camera views.

To mitigate such visual ambiguity and appearance variations, most existing re-id approaches rely on constructing fully supervised machine learning models with extensively labelled training datasets which is unscalable for practical applications in the real-world. Particularly, human annotators must exhaustively search over a vast quantity of offline collected data, manually label cross-view matched images of a large population between every possible camera pair. Nonetheless, having the prohibitively expensive human efforts dissipated, a trained re-id model is often not easily generalisable and transferable, due to the elastic and dynamic operating conditions of a surveillance system. With such motivations, this thesis proposes several scalable re-id approaches with significantly reduced human supervision, readily applied to practical applications. More specifically, this thesis has developed and investigated four new approaches for reducing human labelling effort in real-world re-id as follows:

Chapter 3 The first approach is *affinity mining from unlabelled data*. Different from most existing supervised approaches, this work aims to model the discriminative information for re-id without exploiting human annotations, but from the vast amount of unlabelled person image data, thus applicable to both *semi-supervised* and *unsupervised re-id*. It is non-trivial since the human annotated identity matching correspondence is often the key to discriminative re-id modelling. In this chapter, an alternative strategy is explored by specifically mining two types of affinity relationships among unlabelled data: (1) inter-view data affinity and (2) intra-view data affinity. In particular, with such affinity information encoded as constraints, a Regularised Kernel Subspace Learning model is developed to explicitly reduce inter-view appearance variations and meanwhile enhance intra-view appearance disparity for more discriminative re-id matching. Consequently, annotation costs can be immensely alleviated and a scalable re-id model is readily to be leveraged to plenty of unlabelled data which is inexpensive to collect.

Chapter 4 The second approach is *saliency discovery from unlabelled data*. This chapter continues to investigate the problem of what can be learned in unlabelled images *without* identity

labels annotated by human. Other than affinity mining as proposed by Chapter 3, a different solution is proposed. That is, to discover localised visual appearance saliency of person appearances. Intuitively, salient and atypical appearances of human are able to uniquely and representatively describe and identify an individual, whilst also often robust to view changes and detection variances. Motivated by this, an unsupervised Generative Topic Saliency model is proposed to jointly perform foreground extraction, saliency detection, as well as discriminative re-id matching. This approach completely avoids the exhaustive annotation effort for model training, and thus better scales to real-world applications. Moreover, its automatically discovered re-id saliency representations are shown to be semantically interpretable, suitable for generating useful visual analysis for deployable user-oriented software tools.

Chapter 5 The third approach is *incremental learning from actively labelled data.* Since learning from unlabelled data alone yields less discriminative matching results, and in some cases there will be limited human labelling resources available for re-id modelling, this chapter thus investigate the problem of how to maximise a model's discriminative capability with minimised labelling efforts. The challenges are to (1) automatically select the most representative data from a vast number of noisy/ambiguous unlabelled data in order to maximise model discrimination capacity; and (2) incrementally update the model parameters to accelerate machine responses and reduce human waiting time. To that end, this thesis proposes a regression based re-id model, characterised by its very fast and efficient incremental model updates. Furthermore, an effective active data sampling algorithm with three novel joint exploration-exploitation criteria is designed, to make automatic data selection feasible with notably reduced human labelling costs. Such an approach ensures annotations to be spent only on *very few* data samples which are most critical to model's generalisation capability, instead of being exhausted by blindly labelling many noisy and redundant training samples.

Chapter 6 The last technical area of this thesis is *human-in-the-loop learning from relevance feedback.* Whilst former chapters mainly investigate techniques to reduce human supervision for model training, this chapter motivates a novel research area to further minimise human efforts spent in the re-id deployment stage. In real-world applications where camera network and potential gallery size increases dramatically, even the state-of-the-art re-id models generate much inferior re-id performances and human involvements at deployment stage is inevitable. To minimise such human efforts and maximise re-id performance, this thesis explores an alternative approach to re-id by formulating a hybrid human-computer learning paradigm with humans in the model matching loop. Specifically, a Human Verification Incremental Learning model is formulated which does not require any pre-labelled training data, therefore scalable to new camera pairs; Moreover, the proposed model learns cumulatively from human feedback to provide an instant improvement to re-id ranking of each probe on-the-fly, thus scalable to large gallery sizes. It has been demonstrated that the proposed re-id model achieves significantly superior re-id results whilst only consumes much less human supervision effort.

For facilitating a holistic understanding about this thesis, the main studies are summarised and framed into a graphical abstract as shown in Figure 1.

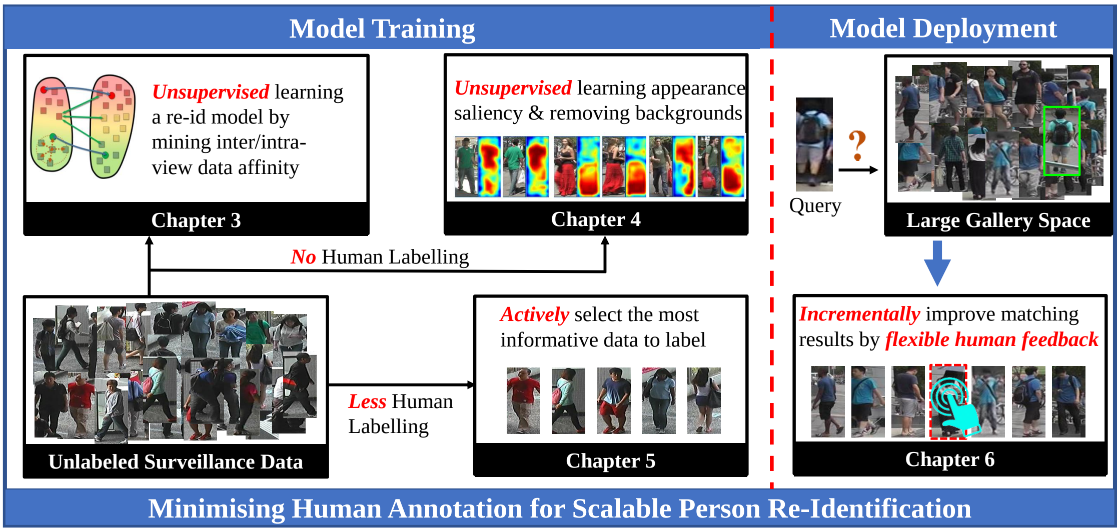

Model Training
Model Deployment
Unsupervised learning a re-id model by mining inter/intra-view data affinity
Chapter 3
Unsupervised learning appearance saliency & removing backgrounds
Chapter 4
No Human Labelling
Unlabeled Surveillance Data
Less Human Labelling
Actively select the most informative data to label
Chapter 5
Query
?
Large Gallery Space
Incrementally improve matching results by flexible human feedback
Chapter 6
Minimising Human Annotation for Scalable Person Re-Identification

Figure 1: An overview of the main studies carried out in this thesis.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged.

Some parts of the work have previously been published or in submission as:

Chapter 3

- H. Wang, X. Zhu, T. Xiang, S. Gong, *Towards Unsupervised Open-Set Person Re-Identification*, IEEE International Conference on Image Processing (ICIP), 2016.

Chapter 4

- H. Wang, S. Gong, T. Xiang, *Unsupervised Learning of Generative Topic Saliency for Person Re-Identification*, British Machine Vision Conference (BMVC), 2014

Chapter 5

- H.Wang, S.Gong, T.Xiang, *Highly Efficient Regression for Scalable Person Re-Identification*, British Machine Vision Conference (BMVC), 2016.
- H.Wang, X.Zhu, S.Gong, T.Xiang, *Person Re-Identification in Identity Regression Space*, submitted to International Journal of Computer Vision (IJCV), 2017.

Chapter 6

- H. Wang, S. Gong, X. Zhu, T. Xiang, *Human-In-The-Loop Person Re-Identification*, European Conference on Computer Vision (ECCV), 2016.
- H. Wang, S. Gong, X. Zhu, T. Xiang, *Human-In-The-Loop Person Re-Identification*, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017.

Acknowledgements

Foremost, I would like to express my greatest gratitude to my first supervisor, Prof. Shaogang Gong, and my second supervisor, Dr. Tao Xiang, for their consistent support and encouragement, as well as all those bright ideas and wonderful opportunities they have kindly provided to me. I consider my self an extremely lucky one to have two of such outstanding supervisors throughout my PhD study. I would like to convey my deep gratefulness to Dr. Xiatian Zhu, not only because of the countless discussions, emails, chats, late working nights we went through together, but also that he was always there to fill me up with faith in my work when I was confronted by failures. It was through their guidance that I gradually became an independent researcher.

I wish to thank Dr. Yi-Zhe Song for introducing me into the group at the first place back in 2013, and Dr. Tim Hospedales for being my independent assessor throughout my PhD study. I would also like to thank my roommates who later became my best friends, Dr. Xingyu Han and Dr. Zhiyuan Shi, for all the laughs and happy times we have experienced together. My warm appreciation goes to everyone I met at the Vision Group for their friendship and support: Ke Chen, Yanwei Fu, Ryan Layne, Brais Cancela, Yi Li, Xun Xu, Li Zhang, Yun Zhou, Yongxin Yang, Kunkun Pang, Ioannis Alexiou, Heng Yang, Elyor Kodirov, Taiqing Wang, Xiaolong Ma, Yaowei Wang, Xiangyu Kong, Shuxin Ouyang, Jingya Wang, Qian Yu, Feng Liu, Arne Schumann, Ying Zhang, Qi Dong, Xiaobin Chang, Jifei Song, Hang Su, Da Li, Wei Li, Xu Lan, Zhiyi Cheng, Kaiyue Pang, Conghui Hu, Yanbei Chen, Umar Riaz Muhammad. I am very grateful to the EECS administrative staff and system supports staff, especially Mellisa Yeo and Tim Kay for taking care of my progresses in PhD study as well as in occupying school computing servers.

My deep and sincere gratitude to my parents for their continuous and unconditional love. I owe them too much. Last but not the least, this journey would not have been possible without the love and devotion from my loving fiance Xinhui Hu. Every journey of hers was flying over five thousands of miles to be accompanied with me, but what I received from her was only endless support without any complaints.

Contents

List of Figures

List of Acronyms

Re-Id	Re-Identification
CCTV	Closed-Circuit Television
CUHK	Chinese University of Hong Kong (dataset)
VIPeR	Viewpoint Invariant Pedestrian Recognition (dataset)
iLIDS	Imagery Library for Intelligent Detection Systems (dataset)
CNN	Convolutional Neural Network
SSS	Small Sample Size
SVM	Support Vector Machine
CCA	Canonical Correlation Analysis
HIL	Human-In-the-Loop
HOL	Human-Out-of-the-Loop
POP	Post-rank Optimisation
OS^2	One-Shot Open-Set
RKSL	Regularsied Kernel Subspace Learning
RKHS	Reproducing Kernel Hilbert Space
ROC	Receiver Operating Characteristic

CMC	Cumulative Match Characteristic
ER	Expected Rank
mAP	Mean Average Precision
R1	Rank-1 Recognition Rage
FAR	False Accept Rate
DIR	Detection and Identification Rate
GTS	Generative Topic Saliency
IRS	Identity Regression Space
FDA	Fisher Discriminant Analysis
HVIL	Human Verification Incremental Learning
RMEL	Regularised Metric Ensemble Learning
SGD	Stochastic Gradient Descent
LEGO	LogDet Exact Gradient Online

Chapter 1

Introduction

1.1 Person Re-Identification in Surveillance

1.1.1 Motivation

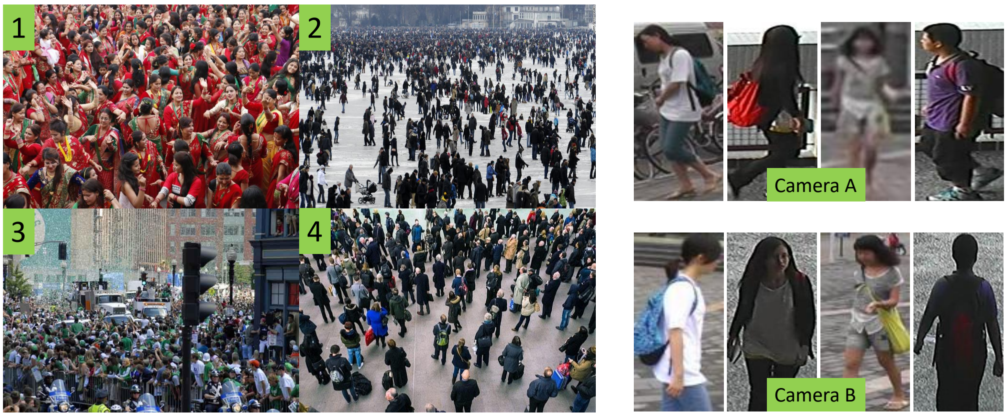
Person Re-Identification (re-id) refers to the problem of visually matching an individual or a group of people across non-overlapping cameras distributed at diverse physical locations and times [3]. For most of today's intelligent surveillance systems, re-identification has become a fundamental functionality which paves the way for numerous higher level and more complex applications. For example, it contributes as a critical component for a multi-camera tracking or forensic search system, which allow government agencies to fast locate suspicious criminals, and therefore prevent terrorism threatening social infrastructure and civilian safety and security; The re-identification of a group of people collectively provides valuable intelligence for crowd movement/behaviour analysis, which facilitates public spaces like airports or shopping malls to conduct better crowd control practices or develop more profitable retail floor plans; Re-identification techniques could also be integrated into smart home automation platforms, so as to enable functionalities such as elderly/baby monitoring, intrusion detection and burglary alarming.

Among the various fields to which re-identification technologies could bring benefits, the most significant application scenario is the one encountered by visual surveillance systems operating over large closed-circuit television (CCTV) camera networks. Thanks to the technical innovations and the availability of cheaper and more advanced electrical equipments in the past decades, the deployments of CCTV networks are fast-growing and wide-spreading, currently prevalent in public spaces of every major city worldwide. By the year of 2016, there are about 350 million CCTV cameras installed globally [4]; It was estimated by [5] that Britain has 1 surveillance camera for every 11 people in UK. According to a recent report from MarketsandMarkets [6], the video surveillance market was valued at USD 30.37 Billion in 2016 and is projected to reach USD 75.64 Billion by 2022, at a CAGR (Compound Annual Growth Rate) of 15.4% between 2017 and 2022. Willing or not, most of contemporary human beings have already become permanent residuals of a surveillance state, and meanwhile been benefiting from the mass convenience and value brought by the mass surveillance.

1.1.2 Recent Developments

However, re-identification is never trivial in these real-world scenes at large scales, and still remains unsolved to both academic and industrial communities. In particular, for surveillance systems in the crowded and unconstrained public spaces, re-identification relying upon higher-level biometry such as face recognition is neither feasible nor reliable, due to numerous complicated factors such as uncontrolled standoff distances (distances between the camera and the object), insufficient image details, low camera resolution, and so forth. Instead, researchers have turned to alternative solutions by exploiting the holistic appearances of people, whose visual features predominated by their clothing, skin color and objects carried or associated with them. However, such physical characteristics are intrinsically weaker and consequently guarantee much lower identity matching accuracies. For instance, many people may dress alike due to cultural traditions, locations, seasons, social norms in fashion and so on, every of which undermining the discrimination capability of this kind of representation. Moreover, what further compounds the problem is that person visual appearances may undergo dramatic variations in different camera views caused by the unconstrained viewing condition, e.g. illumination, occlusion, background clutter, and human pose. In other words, a re-identification system is required to differentiate person images often with high intra-class variances and low inter-class variances (Figure 1.1).

In order to address such problems, the predominant approaches in existing literature follow a standard supervised learning scenario. That is, to exploit manually labelling by human experts as externally provided information. For instance, the human labelling could specify the identity information of each individual; Or it could specify whether a given paired person images captured by two different cameras belong to the same identity or not, regardless their visual appearance dissimilarity/similarity. Trained with such labels, a machine learning model is



(a) Different people have similar appearances.

(b) Cross-view appearance variations

Figure 1.1: (a) Examples of low inter-class variances in people appearances: (a1) Women in Hinduism wearing red; (a2) People in winter wearing dark; (a3) Sport fans wearing team colors; (a4) Workers wearing suits. (b) Examples of high inter-class variances in people appearances: The same four individuals observed by two different cameras (Each column indicates one identity).

therefore more capable of discriminating person identities, more sensitive to subtle differences in person appearances and more robust to viewing condition variations. Such a trained model can be then leveraged for automatically matching person identities during deployment stage. It is evident that the external annotations from human experts, i.e. the training labels, participate as one indispensable component in the procedure of knowledge transferring from human expertise to an automated re-identification model. Based on this supervised learning approach, the re-identification community has witnessed ever-increased matching accuracies on increasingly larger sized benchmarks of more training identity classes over the past two years. For instance, the CUHK03 benchmark [1] contains 13,164 images of 1,360 identities, of which 1,260 are used for training with 100 for testing, significantly larger than the earlier VIPeR [7] (1,264 images of 632 people with 316 for training), and iLIDS [8] (476 images for 119 people with 69 for training). The state-of-the-art Rank-1 accuracy on CUHK03 has exceeded 80% [9], tripling the best performance reported only two years ago [1] (Figure 1.2).

Despite such rapid progress, we found these automatic re-identification solutions ill-suited and unscalable for practical deployments due to human labelling. More specifically, these methods are based on a few assumptions which are too artificial and unrealistic about human labelling, failed to meet numerous real-world challenges. They will be discussed in the following section.

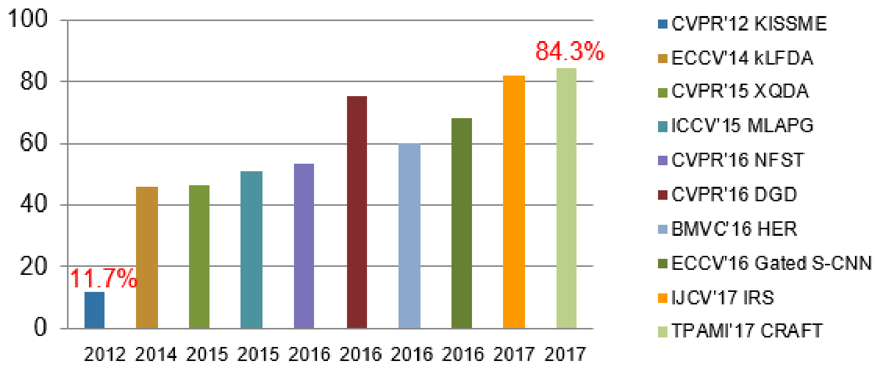


Figure 1.2: Rank-1 recognition rates on CUHK03 [1] published in main conferences.

1.2 Human Labelling for Modelling Re-Identification

An ordinary user can be highly impressed by the intelligence of a deep convolutional neural network [10, 11] (CNN) for its capability of accurately recognising a Welsh Corgi dog in an image (Figure 1.3), but what is often beyond his/her realisation is that the neural network can be trained with datasets containing millions of images labelled by human, and perhaps have witnessed thousands of instances of the Welsh Corgi. Taking the well-known ImageNet [12] dataset as an example, it consists of a total number of 14,197,122 images, each being labelled as at least one of the 21,841 synsets (hierarchical category labels). The most popular category, 'animal', contains over 2799K human labelled images. Such a large scale of training image dataset with accurate human labels is one important factor, if not the most, for recent computer algorithms to successfully conquer many vision tasks, such as image classification [11], object detection [13], segmentation [14], and so forth. Human labelling has become so important that it is common to see many companies such as IBM hiring labelers or outsourcing the labelling work through online platforms such as Amazon Mechanical Turk [15].

1.2.1 Challenges to Human Labelling

However, compared to most above listed vision tasks, there exists many more difficulties and challenges to exploit human labelling for re-identification. Specifically, the challenges to human labelling for re-identification in real-world applications are summarised as following.

1. ***Labelling Cost***: Identifying and labelling person identities in a large scale of CCTV surveillance videos is intrinsically harder and more expensive compared to other more common annotation tasks in vision such assigning image class labels (classification), or

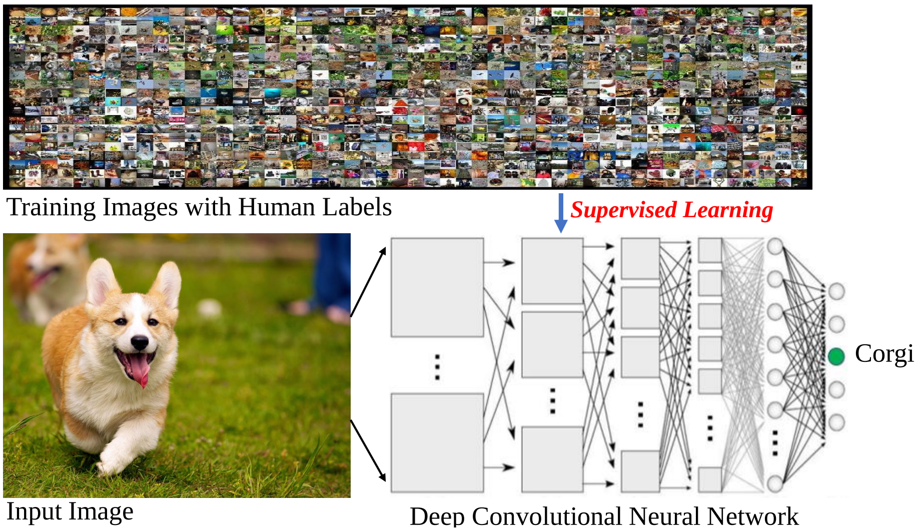

Training Images with Human Labels
Supervised Learning
Corgi
Input Image
Deep Convolutional Neural Network

Figure 1.3: Image classification by a fully-supervised deep neural network.

drawing bounding boxes of objects (localisation/detection). Specifically, labelling the identity of a person requires several decision processes: *Does this person belong to those already labelled identities stored in the database? If yes, which one should it be assign to? If not, should it be assigned a new ID or discarded?* The procedure of telling *'who it is'* is apparently more complicated and tedious than just telling *'what it is'*, i.e. identifying a image by *'This is (not) a person'* as in other more common classification tasks. What further compounds labelling identities are the homogeneous appearance among different persons and the dramatic appearance variations across camera views, i.e. small inter-person variances and large intra-person variances, demanding more energy, time, and concentration of a human annotator. For instance, [16] reported the average work shift of a modern CCTV operator was now 12-hours, which is a much heavier work overload compared to the average working time.

2. ***Expertise Requirement***: The human expertise required by the labelling person identities in surveillance camera networks is substantially high. Generating accurate and efficient human labelling for re-identification requires years of professional work experience. More specifically, a human operator needs to be capable of correctly infer person identities from surveillance video frames whose contents are noisy and cluttered, often with low image resolutions and large numbers of candidates per frame. Furthermore, the operator needs to have sufficient context knowledge, i.e. being familiar with the physical layout of the camera network(s), and the frequent trajectory choices of a pedestrian, so as to predict and

search over the cameras and time frames where the target will most possibly to re-appear. Such a labelling job can hardly be outsourced through online mechanical platforms to workers without professional training, and it is difficult to transfer this expertise directly between operators. As a result, the difficulty to obtain sufficient labels further increases.

3. ***Plausibility***: One fundamental challenge is that, to obtain 'sufficient' labels for person re-id might just be not plausible. Particularly, due to the uncontrolled pedestrian path and open-ended environment, there simply may not exist enough person identities who re-appear in *every* different cameras of a public surveillance network. In fact, the labelled training population for re-identification is often small in number, e.g. hundreds of person classes, and the training samples of each person class is also often limited, in some cases only one-shot of the person being available. The training sample size is thus much smaller (often in an order of magnitude or more) than the typical feature dimension. This lack of training samples is known as the Small Sample Size (SSS) problem [17]. The SSS problem can thus result in ill-estimated intra-class variances, indications of problematic class distributions, which in turn lead to suboptimal discriminative solutions.

4. ***Scalability***: Even if there are sufficient people who indeed re-appear in all camera views, to obtain a manually pre-labelled pairwise training data set *for every camera pair* requires continuous monitoring and exhaustive searching which is infeasible and unscalable in practice. In a real-world topologically complex and large camera network, there are a quadratic number of camera pairs with a extremely large search space for labelling. Such a *scalability* challenge is another cause for the aforementioned SSS problem.

5. ***Generalisation***: A trained re-identification model with samples labelled in one specific camera network usually cannot generalise or transfer perfectly to other camera networks with different camera configurations (e.g. image resolution, camera focus), viewing conditions (e.g. viewing angle, illumination, background clutter), physical topologies, etc. In other words, the re-identification labelling is usually camera-network specific, constrained by many factors related to one particular network and thus difficult to generalise to others.

6. ***Adaptability***: Even for the same camera network, the operating condition also varies over time instead of being a constant factor. For instance, the illumination may change at different times of day; The viewing background may change due to different weathers (raining,

snowing, cloudy, etc); The population appearance pattern may vary in different seasons, or evolve over years. As a result, the human labelling obtained at a specific time period may not be adaptable to the elastic viewing conditions. New human labelling with extra costs will be needed again to update the re-identification model.

1.2.2 Hypotheses of Existing Approaches

Given all the listed challenges, one inevitable question arises: How well do state-of-the-art re-id approaches tackle these challenges? Unfortunately, most of existing re-id works fail to take any of such challenges into consideration in model design, and thus are still far from an automated re-id solution capable of deployment in the real-world. In particular, in most state-of-the-art methods [1, 18, 19, 20, 21, 22, 23, 24, 25], a re-id model is trained with an overwhelming demand and heavy reliance on a large scale of human labels, without taking into account the expense, feasibility and other challenges of real-world label collection. Particularly, they usually share four common artificial hypotheses:

1. ***Closed-world matching***: Many approaches assume that a re-identification model works in a extremely constrained scenario where a person in one camera *must* re-appears in other cameras. In the model training phase, this hypothesis is reflected by the fact that every training person identity is annotated under at least one pair of camera views, so that his/her cross-view appearance variations are guaranteed to be labelled. It is evident that this hypothesis largely underestimates the *labelling cost*, *expertise requirement*, and *plausibility* for real-world human labelling. Moreover, this hypothesis has also been reflected in unrealistic testing evaluations. For instance, most existing works test a re-id model by matching two sets of person images, namely the probe set and gallery set, which contain exactly the same group of people. In other words, every testing person in one set is guaranteed with prior knowledge to be definitely included in the other set. This is however another strong closed-world assumption. In practical environments, the person identities of the probe and gallery set could be only partially overlapped, and there exist much more distractors in the potential searching space than the target persons.

2. ***Offline training label collection***: Existing supervised learning based re-id approaches artificially assumes an offline training label collection process. That is, a pre-labelled training dataset containing either binary-class labelled true/false-matching image pairs or multi-

class labelled individual person images is collected by human annotators for every pair of cameras through manually examining a vast pool of image/video data. This training dataset is then used to train an offline re-id model. However, due to the aforementioned *generalisation* and *adaptability* challenges, it is highly possible that such an offline trained re-id model will not generalise/adapt perfectly to various camera networks, viewing conditions and population appearance patterns which vary over time. In fact, real-world data collection and model training is more incremental than static, that is, additional labelled images are generated over time and available for new model training. It is thus highly desirable for a re-id model to incorporate increasingly available labelled data, growing and adapting continuously to the changing environments.

3. ***Small testing population***: In most popular re-id benchmark datasets [1, 7, 8], the size of the training population is either significantly greater or no less than that of the testing population. For instance, the standard CUHK03 benchmark test defines the training set having paired images of 1,260 people from six different camera views (on average 4.8 image samples per person per camera view), whilst the test set having only 100 identities each with a single image. The test population is thus 10 times smaller than the training population, with approximately 50 times less images. This is however another erroneous experiment design led by an unrealistic assumption. In practice, any deployment gallery size (test population) is almost always much greater than any labelled training data size even if such training data were available. In a public space such as an underground station, there are easily thousands of people passing through a camera every hour [26] with a testing gallery population size of over 10,000 per day, much more than the amount of affordable human labels for training.

4. ***Fully-automated deployment***: The above discussed *offline training label collection* and *small testing population* assumptions often resulted in an blind confidence on fully-automated model deployments. It is tacitly assumed by most that an offline-trained re-id model is capable of re-identifying target (unseen during model training) person images at test time in a fully-automated manner, without any human assistance nor model adaptation. As the testing population sizes in most standard re-id benchmarks are small, existing fully-automated approaches have achieved sound matching accuracies and this hypothesis seems unchallenged. However, we observed on CUHK03 dataset that, a 10-fold increase in gallery size

leads to a 10-fold decrease in re-id Rank-1 performance, even when the state-of-the-art re-id models were trained from sufficiently sized labelled data. Given such low Rank-1 scores, in practice human operators (users) would still be required to verify any true match of a probe from an automatically generated ranking list. Consequently, how to efficiently exploit human labelling effort in a cost-effective way during the deployment stage arises as an open question which is however largely overlooked by existing methods.

1.3 Contributions

The research of this thesis attempts to move one step further toward re-identification applications in practice by proposing several re-id models to specifically address the human labelling challenges in the real-world and relax hypotheses which are practically unrealistic. Specifically, the contributions of this thesis to re-identification research are summarised below:

1. Chapter 3: A new subspace learning based re-id model is proposed to exploit inter/intra-view affinity information from *unlabelled* data, with an efficient and flexible solution which can be applied to both *semi-supervised* and *unsupervised* re-id. The capability of learning from *unlabelled data* substantially reduces the demand of heavy human labelling for model training, and completely avoids the human labelling challenges discussed in Section 1.2.1. Furthermore, to relax the unrealistic hypothesis of *closed world matching*, a new *OneShot-OpenSet Re-Id* problem setting is introduced. It poses more realistic challenges to the research community and paves a way towards large scale open-world re-id.

2. Chapter 4: Instead of only learning a general matching function (Chapter 4), in this Chapter a new *unsupervised* re-id model is proposed, aiming to explore more fine-grained image details from the *unlabelled data*. Specifically, a novel generative saliency discovery model is proposed which is capable of simultaneous foreground saliency detection, background clutter removal and re-id matching, without any forms of human labelling. As a completely *unsupervised* approach, it significantly improves the *scalability* of a re-id model. In addition to re-id matching, its automatically discovered foreground saliency is also useful as an image analysis module whose target users are human operators of a surveillance system.

3. Chapter 5: A new active learning algorithm for cost-effective human labelling is proposed to reduce *labelling cost* and increase *scalability*, by only querying the most informative

rather than randomly sampled feedback from a human operator. This active learning model aims to jointly explore the population diversity and discover the class boundary of the up-to-date model. In addition, to relax the *offline training label collection* and *fully-automated deployment* hypotheses, a regression based re-id model is formulated, enabling to rapidly update an incremental re-id model from piecewise new data *only*, and progressively adapt the model to more data when available.

4. Chapter 6: A hybrid human-computer learning paradigm is proposed to minimise the human labelling effort during model deployments. More importantly, a new human-in-the-loop re-id model is formulated with a few advantages: (1) *Scalability*: The model can be directly deployed without the need of heavy human labelling for the pre-collection a separate training dataset. During deployments, it enables a user to re-identify rapidly a given probe person image after only a handful of feedback verifications, without the need for exhaustive eyeball search of true/false in the entire very large gallery set. (2) *Generalisation* and *Adaptability*: It introduces a new online incremental distance metric learning algorithm, which enables the re-id model to cumulatively update parameters to utilise on-the-fly user feedback, and adapt itself to the varying operating conditions.

1.4 Thesis Outline

The remaining chapters of this thesis are organised as follows:

Chapter 2 provides a review of existing research relevant to the main components of this thesis.

Chapter 3 investigates an inter/intra-view affinity mining algorithm to explore discriminative information only from the unlabelled data, so the human labelling can be avoided from training. It also introduces a more realistic open-world re-id setting.

Chapter 4 proposes an unsupervised model which discovers localised saliency regions and removes cluttered backgrounds on person images without the need of human labelling. The discovered salient appearances are shown to be effective in re-id matching.

Chapter 5 presents an active learning based re-id model, which reduces human labelling by selecting only the most informative unlabelled data to actively query. It also considers an incremental learning setting to improve model generalisation and adaptability.

Chapter 6 presents a hybrid human-computer learning paradigm which smooths the boundary of re-id model training and testing. This human-in-the-loop model does not require any labelling for training, and meanwhile exceedingly reduces the human labour spent during deployments. Moreover, it is designed to be updated incrementally from cumulative user feedback, well suited to the real-world scenarios with varying viewing conditions.

Chapter 7 includes concluding remarks and discusses potential areas for future research and extensions.

Chapter 2

Literature Review

2.1 The Re-Identification Problem

The general task of an automated re-identification system is: when being represented with a person of interest, the system needs to tell whether the same person has been observed, and to locate the same identity in the large amount of video footage generated in a network of surveillance cameras watching over public spaces with major pedestrian traffic flows. A standard pipeline [27] for such a person re-id system contains three following modules (Figure 2.1):

1. Pre-processing: This step refers to the generation of images of pedestrians by applying a person detection and tracking process on the raw video frames collected by surveillance cameras. The generated person images are treated as input data for the re-id system.

2. Representation: After the acquisition of person image data, the second stage is to construct a representation of each image or tracklet (a sequence of images), i.e. to extract discriminative visual features to describe individual appearances.

3. Matching: The core module of a re-id system is to match the imagery features of the query (or interchangeably termed as probe) images/tracklets against a gallery of persons by measuring the similarity between features. Often a re-id matching model is required to be trained so that an optimised similarity function can be found.

It is a common belief to the research community that the preprocessing stage, i.e. person detection [28, 29, 13] and tracking [30, 31, 32], should be treated as independent research ar-

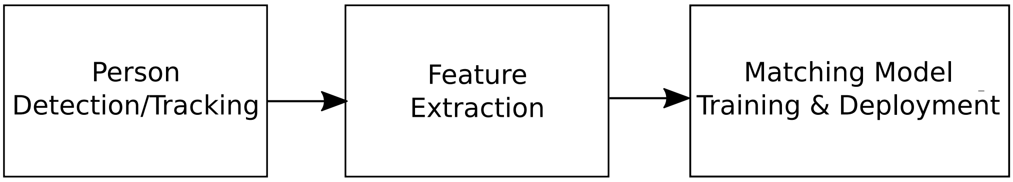

Person Detection/Tracking
Feature Extraction
Matching Model Training & Deployment

Figure 2.1: The pipeline of a person re-identification system.

eas, and therefore interested readers are invited to read aforementioned references for more details of them. This chapter reviews particularly the most recent developments in the other two core stages of re-identification, i.e. different strategy in designing feature representations and learning matching models. Specifically, Section 2.2 reviews a selection of broadly used feature representations for contemporary works; Section 2.3 discusses various re-id model learning and deployment strategies as well as their connections to the contributions of relevant chapters in this thesis.

2.2 Feature Representation

Feature representation is an important step within the re-id process. The choice of feature is critic since it needs be robust to the changing factors like illumination, viewpoint, occlusion and image resolution. In some early works on re-identification, researchers have been exploring hand-engineered low-level features such as color histograms and texture filters to represent human appearances; Later on several mid-level descriptors which are more robust to viewing condition variations were proposed; Most recently, along with the development of powerful deep convolutional neural networks, discriminative representations can also be directly learned with raw image pixels. This section mainly reviews the first two types of features, whilst the last type will be discussed in Section 2.3 together with other model learning strategies, since the learning of deep representations is essentially one type of re-id model.

2.2.1 Low-level Features

There are two types of hand-engineered low-level features being popularly used, reflecting color distributions and edge/texture properties of a region respectively. The features are often represented as bag-of-words scheme in the form of histogram. The color features includes color histograms in different color channels. Color spaces like RGB, HSV, Lab, and YCbCr are often explored. As to the edge/texture feature, existing works often use Scale-invariant feature transform (SIFT) Descriptors [33], local binary patterns [34] or texture filters like the Gabor filter and the Schmid filter [35, 36, 37] to represent texture and gradient information of a given image region.

A natural consideration next is how to divide a whole image into regions. As extracting feature histograms on a whole image would be inaccurate and unreliable to reflect the important information on localised details, existing works often first separate an image into different local regions and then extract features on each region. Because of viewpoint changes and arbitrary pedestrian poses, an individual appearing in a image caught by one camera usually does not appear in the same region within another image caught by a different camera. This problem is known as the mis-alignment problem. To avoid the mis-alignment problem, two types of image segmentation schemes have been proposed: part-based and patch-based representations.

Part-Based Representation It tends to divide the images according to different parts of human body. For example, Gray *et al* [35] and Prosser *et al* [36] divide the whole image into 6 equal sized horizontal strips in order to roughly captured the head, upper and lower torso and upper and lower legs. In this scheme they believe individuals could appear in different positions in different images, but the body parts should remain the same horizontally. After that, they use color features as 8 color channels (RGB, HSV and YCbCr) and 21 texture filters (8 Gabor filters and 13 Schmid filters). Then each feature is represented by a 16-bin histogram. So for each strip, the feature vector is of $(8+21)\times 16 = 464$ dimension. And the final representation of the image is the concatenation of the six strips' features, ending up in a 2784-dimensional feature vector.

Another part-based segmentation method is proposed by Farenzena *et al* [38]. They explore the principles of symmetry and asymmetry, using two horizontal axes of asymmetry that isolate three main body parts (head, torso and legs) and two vertical axes of symmetry to isolate the left part and right part of torso and legs. After dividing an individual's figure into 5 parts, they extract color features as weighted color histograms for each part where pixels near the vertical axes gain more weight. Then they use RHSP (Recurrent High-Structured Patches) to encode edge/texture feature for each part. Also, they introduced MSCR feature (Maximally Stable Color Regions) which represent the information like area and centroid of blobs having a stable color and can also be treated as a color feature.

Other than [38] which only uses some geometry assumptions to roughly locate the body parts, there are also works which explicitly utilize body-part detectors to explore the body configurations, such as the work of [39, 40]. For example, [39] utilizes pictorial structures part detector trained elsewhere on re-id images, and extracts low level features like colour histogram and MSCR within the detected body mask.

Patch-Based Representation Different from part-based representation, patch-based representation divide the images into regular sized local patches aligned in grids for matching persons. For example Zhao *et al* [33] and Hirzer *et al* [34] both took this way for feature extraction. Considering the problem of mis-alignment caused by viewpoint change and pose variation, normally the patches are overlapped, trying to catch slight movement of human body. While Hirzer *et al* [34] concatenate the feature vectors of all the patches to a single feature vector to represent a whole image, Zhao *et al* [33] retain the patch representation and take more steps to handle mis-alignment problem, which will be discussed later.

On feature selection, the work of [34] choose the mean value of patch pixels in HSV and Lab color channels to represent color information, and use Local Binary Patterns (LBP) to catch edge/texture information. The patches are in the size of 8×16 pixels, sampled on a grid of 4×8 pixels. So the patches are 50% overlapped both horizontally and vertically. In the work of [33], they choose color histograms in Lab space as color features and use SIFT descriptor as edge/texture features. The patches are of size 10×10 and sampled on a grid with a grid step size of 4 pixels. So their patches are also overlapped in both direction.

2.2.2 Mid-level Features

Compared to the low-level feature representations above, the mid-level features usually are more effective since they are less vulnerable to varying conditions like illuminations and poses. Most mid-level representation requires some extent of learning, and thus have more discriminative power than hand-engineered features. There are several mid-level representation methods as discussed below:

Semantic attributes are used as mid-level representations for re-id, firstly introduced by Layne *et al* in [41] and [42]. The authors proposed a method that learns a selection and weighting of mid-level semantic attributes to describe people. Different from low-level feature representations and high-level classes/identities, attributes provide a mid-level semantic description of an instance. After the low-level feature extraction, they train a Support Vector Machine (SVM) to detect attributes. Using the SVM, each image can be interpreted as an attribute profile, which reflects the SVM's confidence on each attribute existing on this certain image. This attribute profile can be treated as a new feature representation, and fused as a complementary to low-level features. In Li *et al*'s work of [43], a more complete attribute topology is defined and learned through a latent SVM model.

Mid-level filters are also explored as one type of mid-level representation in the work of [21]. The key idea is that certain patches on one image could be more effective on describing one person because they are neither too rare nor too general throughout the dataset. Thus filter responses on those effective patches can form a good mid-level representation for the task of re-id. Through clustering and supervision, each learned filter response is coherent in appearance, specific for location, and also robust for cross-view variation.

Other than attributes and mid-level filters, Ma *et al* [44] have explored fisher vector based representation for the task of re-id. They combine Fisher vectors with a local descriptor and use the resultant representation (*Local Descriptors encoded by Fisher Vector* or LDFV) to describe persons images. The method also shows promising performance when combined with metric learning approaches. More recently, Yang *et al*'s proposed method [45] explores another mid-level representation - the salient color names. By mapping the raw RGB color space values to a probability distribution over a 16 dimension color names, the proposed representation also gives state-of-art result when combined with supervised metric learning methods. All of the above methods have provided a insight on the potential of mid-level representations to improve re-id results.

2.3 Matching Model

Feature representations alone are often insufficient to accurately capture complex appearance variations across cameras with uncontrolled viewing conditions as typical in visual surveillance scenarios. A matching model is thus needed to obtain a more robust and reliable cross-view image similarity/distance measurement. In this section, re-id matching models are discussed from various aspects including: supervision strategies during model training, model updating strategies when new data becomes available, and the deployment strategies on how human operators interact with the re-id model when it is being leveraged.

2.3.1 Supervision Strategies

Supervised Learning Most existing re-id models are *fully-supervised* learning models, usually framed into classification [9, 18, 20, 21, 23, 24, 25, 46, 47, 48, 49], pairwise verification [1, 50, 51, 52], triplet ranking [22, 36, 53, 54, 55, 56, 57], or their combination [58]. These supervised models require a large amount of exhaustively labelled cross-view matching image pairs for each pair of cameras. Such a heavy and annotation requirement significantly restricts their use in real-world settings, and more importantly, their scalability to large camera networks with many camera pairs.

Many of the above approaches can be formulated as a Mahalanobis metric learning problem. For instance, PCCA [59] learns a projection space with a hinge loss function, and constraints thresholding on the margin over distances between matched image pairs as well as unmatched pairs. Similar approaches can be found in LFDA [20, 23] and PRDC [19], stating that distances between matched pairs should be either strictly minimized or relatively smaller than distances between unmatched pairs. While models listed above only learn one global linear projection matrix (thus treating images from both view equally), supervised multi-view subspace learning methods like Canonical Correlation Analysis(CCA) [60, 61] have also been explored to better handle the modality shift caused by the viewing condition variation.

Inspired by the success of deep learning in other computer vision problems, deep re-id models [1, 9, 52, 51, 57, 62, 63] have recently attracted more attention and made significant progress in improving re-id performance. This trend is mainly driven by the availability of larger re-id datasets such as CUHK03 [1] and Market-1501 [2]. These deep networks often contains millions of parameters, constructed by a stack of convolution layers and fully connected layers to learn discriminative image features, and trained by iterative optimisations on a large amount of labelled training data. However,since these deep learning based methods are data-hungry and require more training data to be labelled, the scalability problem becomes even more acute.

Transfer Learning The scalability limitation of these supervised methods has motivated a number of transfer learning-based methods [64, 65, 66, 67]. These methods aim to extract and employ the transferable knowledge from the labelled data in auxiliary datasets for assisting the learning of the target model. Often, a strong relevance between auxiliary and target datasets is assumed. However, they suffer from the generalisation problem. In particular, the difficulties in extracting domain-invariant knowledge and the significant unknown viewing condition variations and often yield ineffective re-id models. In addition, they still bear the assumption that sufficient labelled information is available and needs to be labelled in the source domain.

Unsupervised Learning Unsupervised methods do not require labelled image pairs, and thus are able to scale up to large surveillance camera networks in real-world. However, very few unsupervised methods exist, since it is much harder than supervised learning from labelled information on person-specific appearance. Earlier unsupervised learning re-id methods are focused on feature design [38, 44, 68]. Later on, Liu et al. [37] proposed a feature importance mining scheme, aiming to optimise the weights for global feature types. Nonetheless, their re-id matching performance is less appealing, since it is very hard to design or select effective identity-discriminative features, due to the unknown large cross-view covariates. Zhao et al. [33] proposed a patch-based representation to learn local saliency in a person's appearance which are shown to be effective for re-id matching. However, this approach is exhaustively data-driven therefore computationally complex. This is due to the fact that the approach is based on constructing a different saliency model for every local image patch in every image against a reference set whilst each image is decomposed into hundreds of patches. That is, if there are M images to be matched across two camera views and each image is decomposed to N patches, there are $M \times N$ different saliency models required to be constructed against the reference set. This data-driven approach to unsupervised saliency learning also makes it potentially unstable to large scale problems. For these problems, many images of people (from hundreds to thousands) need be matched across camera views and peoples appearance necessarily exhibits greater variety.

Compared to these existing methods, the two unsupervised methods proposed in Chapter 3 and Chapter 4 improve significantly in both matching accuracies as well as computational efficiency: In particular, Chapter 3 exploits the soft-correspondences across camera views to compensate for the lack of manually labelled cross-view data pairs, significantly different from the existing approaches. The problem is framed into a subspace learning model which has a efficiently solved closed-form solution; Chapter 4 improves the saliency detection framework by learning a *single* generative model for computing saliency map for all the images in a camera view, without the need to perform model retraining, significantly reducing model complexity. Moreover, the model segments simultaneously foreground and background, giving more accurate saliency detection compared to [33] as the latter is sensitive to false saliency detection caused by confusing background as salient foreground.

Semi-supervised Learning Lying somewhere in-between supervised learning and unsupervised learning are semi-supervised learning approaches. Semi-supervised models still require some data labels to build optimisation constraints, but they are also able to exploit unlabelled data as constraints for regularising model learning. Few existing works in re-id has explored this area [69, 70]. The work in [69] models the data distribution by exploring the manifold structure of the unlabelled gallery images. Such manifold structure are then explored to propagate some sparse user-labelled samples to the large quantity of unlabelled gallery set. Liu [70] utilizes unlabelled images from each camera view to build better coupled dictionaries for a image patch representation. However, even less labelling is required, these methods still assume the availability of some labelled data. Moreover, for both work the unlabelled data are only exploited independently in each camera view. They make no attempts with the unlabelled data to learn cross-view identity-discriminative information which is critical for matching people across views. Compared to them, the approach proposed in Chapter 3 exploit the cross-view affinity graphs of unlabelled data to specifically capture cross-view identity-discriminative information, and it does not require necessarily the availability of any labels.

Active Learning One possible solution to the scalability problem associated with human labels is to explore active learning techniques. Active learning is a canonical strategy for reducing human labelling effort by selecting most informative and valuable samples to annotate [71, 72]. Two typical scenarios are stream-based [73] and pool-based [74] active learning. For the former, an unlabelled data sample is drawn once at a time from an input source, and the learner needs to decide whether to query or discard it. Whilst the later assumes a large set of pre-collected unlabelled data is available, and often a small set of labelled data also exists for model initialisation. One of the most important elements in active learning is the query selection criterion. Notable schemes of selecting queries include uncertainty sampling (e.g. focusing on model-confusing unlabelled samples since confident ones are more likely to be correct and offer less information) [75], query by committee (e.g. the disagreement based methods that use a committee of hypotheses/models) [73, 76], expected error reduction (e.g. to reduce the expected total number of incorrect predictions) [77]. While the overwhelming majority of existing active learning researches are spent on generic object / scene classification [71, 78, 79, 80, 81, 82, 83, 84], very little attempt has been made for person re-id.

To our knowledge, there exist only two works closely related to our research reported in Chapter 5 an active person identification method [85] and a temporal adaptation based re-id model [86]. Specifically, instead of learning a generalised cross-view matching function, [85] trains multi-class SVM person classifiers on known identities with the final model unable to be deployed to re-identify previously unseen people (i.e. new classes). In other words, the learned model has no generalisation ability as required by person re-id. In addition, this model cannot perform incremental learning as efficiently our proposed method in Chapter 5, since their model update requires expensive re-training from scratch and less suitable for human-in-the-loop like active selection. Martinel et al. [86] explore similarly the active learning idea for incremental re-id model update. In comparison, the active learning algorithm proposed in Chapter 5 is more extensive and comprehensive (i.e. joint exploitation-exploration vs. exploitation alone) with lower computational cost (i.e. no need for iterative optimisation and graph based data clustering) thus more suitable for human-in-the-loop driven incremental re-id model learning.

2.3.2 Updating Strategies

Batch Learning Almost all of contemporary re-id models assume a batch-mode learning scheme, that the training images is made available all at the same time as a single data pool so that an offline re-id model can be trained. However, it is difficult to make these batch-mode approaches adaptable to a surveillance camera network with changing viewing conditions and new data being continuously generated. In particular, for these existing batch-mode methods to incorporate any new data, a system has to keep all the past training data, add the new data as a enlarged data pool, and re-train a new model from scratch. This re-training approach makes them unscalable to large-scale deployment in the real-world.

Incremental Learning Incremental learning concerns the problem of model training from data streams where samples arrive in sequence [87, 88]. As opposite to batch-wise model learning where all training data are assumed already available before (off-line) model training, incremental learning often requires additionally immediate on-line model update for making the model ready to accept new data at any time if possible. In computer vision, incremental learning has been explored in many different tasks, such as image classification [89, 90, 91, 92], object detection [93], and visual tracking [94].

In re-identification, incremental learning is of more practical importance since it enables an re-id model to be adapted to the varying viewing conditions in the long term without the expensive data storage and model re-training. Moreover, useful feedback could be generated by human operators as a re-id system is being deployed. Incremental learning models are able to cumulatively utilise these user feedback to improve the matching accuracy, whereas offline trained models cannot. However, very few incremental learning models have been proposed for re-identification, as reviewed below.

[69] consider optimising the time-consuming and error-prone post-rank visual search stage by formulating a Post-rank OPtimisation (POP) model that aims to refine quickly the ranking lists. This is achieved by incrementally learning a specific model for each probe person from a few number of human selections during the re-identification process. However, by design the POP model is inherently restricted and unscalable due to the need for human feedback on all probe images and the independence nature between individual person-specific models that unfavourably prevents the cumulative benefit of historical human selections upon future person matching and feedback. [86] perform incremental update of a learned re-id model during the deployment phase for maintaining continuously model performance over time. Both approaches require multiple iterations of optimisation to conduct each step of an incremental update, which is time-consuming to an end-user of the system. While sharing a similar spirit in incremental modelling, the incremental models proposed in Chapter 5 and Chapter 6 are uniquely characterised with more efficient optimisation (i.e. a closed-form solution without the need for iterative optimisation or solving eigen-problem).

2.3.3 Deployment Strategies

In general, almost all existing methods are aimed for automated human-out-of-the-loop (HOL) re-id deployment, thus suffering from dramatic performance degradation given a small size training population and a potentially large searching space in practice, even with the best state-of-the-art supervised method [23, 24, 25, 47, 95, 96]. In contrast, Chapter 6 proposes a human-in-the-loop (HIL) re-id deployment framework. The proposed model learns interactively from human online feedback equivalent to a smaller number of *selective* labelling of negative-pair data on-the-fly, therefore costing less human "labelling effort". This section reviews the concept of general interactive learning, as well as contemporary re-identification work which also consider human's active participation during deployment.

Interactive Learning Interactive model learning with human-in-the-loop is attractive for two reasons: (1) It provides a user with tools that can significantly alleviate or even eliminate the need for careful preparation of large-sized training data. (2) It allows to reduce the human labelling effort by exploiting a model's capacity interactively. Human-computer interactive models have been considered in image segmentation [97, 98], object recognition [99, 100], semi-supervised clustering [101] and object counting [102]. In addition, relevance feedback [103, 104, 105] and active learning [106, 80] are also related to a similar idea of exploiting human feedback to improve model learning. The former has been exploited for interactive image retrieval where human feedback to search results are used to refine a query. The latter aims to reduce the human labelling effort by active sample selection for model training. In active learning, knowledge cumulation during model deployment is not considered, and some offline pre-labelled data are typically needed for model initialisation.

Human-In-the-Loop (HIL) Re-Id A small number of HIL re-id methods have been proposed recently. Abir et al. [85] (Fig. 6.2(b,c)) exploited human-in-the-loop verification to expand their multi-class based re-id model. Compared to the approach proposed in Chapter 6, their method requires a pre-labelled training set for model initialisation. Another limitation is that such a model cannot generalise to new person classes re-id when human effort becomes unavailable. Hirzer et al. [107] (Fig. 6.2(d)) considered a form of human feedback which is ill-posed in practice: It only allows a user to verify whether a *true* match is within the top-*k* ranking list. This limits significantly the effectiveness of human feedback and can waste expensive human labour when a true match cannot be found in the top-*k* ranks, which is rather typical for a re-id model trained by small-sized training data and deployed to a larger-size test gallery population. More recently, Liu et al. [69] proposed the POP model (Fig. 6.1(d) and Fig. 6.2(d)), which allows a user to identify correct matches more rapidly and accurately by accommodating more flexible human feedback. However, POP requires to perform label propagation on an affinity graph over all gallery samples. This makes it poor for large gallery sizes (Section 6.5). Moreover, all existing HIL re-id models [69, 85, 107] do not benefit from cumulative learning, i.e. they treat each probe re-id as an independent modelling or retrieval task; therefore the process of model learning for re-id each probe does not benefit learning the models for other probes. This lack of improving model-learning cumulatively from increased human feedback is both suboptimal and disengaging the human in the loop. In contrast, the proposed re-id framework in Chapter 6 (Fig. 6.1(c) and Fig. 6.2(d)) enables incremental model improvement from cumulative human feedback thus maximising and encouraging human-machine interaction.

Chapter 3

Affinity Mining from Unlabelled Data

3.1 Overview

Most existing person re-identification methods assume the availability of extensively labelled cross-view image pairs. However, compared to the small amount of labelled portion, the scale of unlabelled images are much larger and they are also easier to collect with negligible costs. Moreover, most methods assume a closed-world/set matching scenario , i.e. all the probe people exist in the gallery set, and every selected person image are guaranteed to find its cross-view matching pair (see also Section 1.2). These two assumptions significantly limit their usefulness in real-world applications, particularly with large scale camera networks. To relax these assumptions, this chapter focusses on addressing the following two problems: (1) Instead of relying on human annotated data, how to train a discriminative re-id model directly with unlabelled data samples themselves? (2) How to perform re-id in an open-world scenario where the probe population and gallery population are only partially overlapped?

In this chapter, we introduce a new re-id scenario termed *OneShot-OpenSet Re-Id* (OS^{2}Re-Id). Under this setting, there is no assumption on the access to labelled matching pairs, and the probe people are not guaranteed to have a match in the galley set. For re-id under this more challenging yet realistic setting, we propose a novel Regularised Kernel Subspace Learning (RKSL) model. Our RKSL model differs significantly from existing re-id models in its ability to effectively learn cross-view identity-discriminative information from unlabelled data alone, as well as its flexibility of naturally accommodating pairwise labels if available. We demonstrate the

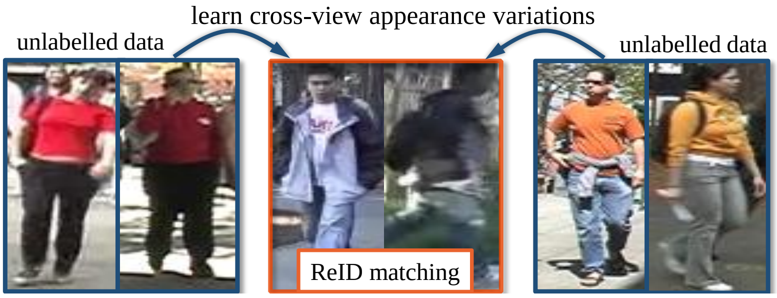


Figure 3.1: Intuition of our cross-view constraint. The unlabelled cross-view data (the left and right pairs) encode information on cross-view appearance variations, e.g. changes in illumination and viewpoint respectively. This subtle information is exploited effectively by the proposed RKSL model for re-identifying the truly matched cross-view people (the middle pair).

efficacy of the proposed model by extensive comparisons with related state-of-the-art methods on two benchmark re-id datasets, VIPeR and CUHK01.

3.2 Problem Definition

Automated person re-identification is an essential yet challenging task due to the rapid expansion of large scale camera networks across our physical world [27]. In a public space monitored by a network of surveillance cameras, person re-id aims to match people across (non-overlapping) camera views. Even in a space of moderate size (e.g. an underground station), there could easily be hundreds or even thousands of people passing through within an hour. In a real-world application scenario, the objective is not to match each and every one. Instead, one typically has a small watch list, which could be a list of known active shoplifters for a shopping mall, or a No Fly List for an airport. An automated re-id system is used to assist human in searching for the people on the watch list from a large volume of video footages. This is an extremely challenging task because a person's appearance can change dramatically due to changes in illumination, view angle, background clutter and occlusion in different camera views. In addition, many of the innocent passers-by may look fairly similar to the people on the watch list. To further compound the problem, there may be only a single shot for each person on the watch list offering insufficient data to learn the appearance variations. We call re-id under this real-world setting the *OneShot-OpenSet Re-Id* (OS^{2}Re-Id) problem.

The objective of this study is to solve this *OS^{2}Re-Id* problem without any labelled inter-camera pairs in order to move one step closer towards large scale person re-identification. To this end, we propose a novel Regularised Kernel Subspace Learning (RKSL) model, which is capable of automatically learning more effectively person identify-discriminative information from unlabelled data, the *only* available data in this new problem setting. The model aims to learn a shared kernalised subspace where after being projected, the probe and gallery data become easier to match than in their original feature space. Such a subspace is learned by constraints on two types of affinity information among unlabelled data samples: (1) affinities between gallery and probe images, regardless of their identities, need to be preserved in the learned subspace; and (2) affinities of visually similar person images from the gallery set need to be separated in the subspace. These two constraints are incorporated as regularisation terms in our subspace learning formulation. Importantly, our model has a closed-form solution which runs efficiently making it suitable for large scale and real-time applications. Furthermore, the model is flexible in that it can be readily extended to exploit pairwise labels when available.

Contributions – Our contributions are: (1) We introduce a new and more realistic person re-identification problem called *OneShot-OpenSet Re-Id* (OS^{2}Re-Id). This problem differs significantly to the existing closed-world Re-Id problem and does not require the tedious exhaustive pairwise labelling. This new re-id problem poses more realistic challenges to the re-id research community and paves a way towards large scale open-world re-id. (2) We present a solution to the OS^{2}Re-Id problem by proposing a new Regularised Kernel Subspace Learning (RKSL) model to exploit the unlabelled data, which can be solved efficiently. (3) We further extend our RKSL model to accommodate any sparse labelled data if available. The efficacy of the proposed RKSL model is extensively evaluated on two of the largest benchmarking re-id datasets (CUHK01 [108] and VIPeR [7]) by extensively comparing with a wide range of relevant state-of-the-art methods including three unsupervised models (SDC [33], SDALF [38], and DASA [109]), one semi-supervised models (SSCDL [70]), and four fully supervised models (RankSVM [36], KISSME [18], kLFDA [23], and KCCA [61]).

3.3 Inter/Intra-View Affinity Mining for Open World Re-Identification

Let us first formally define the *OneShot-OpenSet Re-Id* problem before introducing our proposed model. Suppose we only have unlabelled (in a pairwise inter-camera sense) images of people, including a one-shot watch list of target people G (gallery) seen in camera view X and a larger pool of probe people P from camera view Y. Given a probe image in Y, the objective is to determine (a) whether it matches anyone in the gallery set, and (b) if yes, which one. Note that

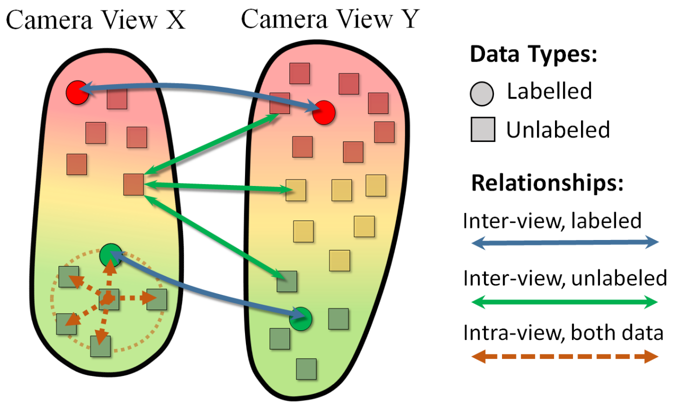


Figure 3.2: Three types of pairwise relationships in re-id. Each node represents a person. Note that View Y caught more people than View X, reflecting the open-set re-id setting. Nodes of the same colour within each view indicate that they have similar visual appearances.

we focus on two views here but an arbitrary number of views can be considered.

Our solution to this OS^{2}Re-Id problem is a *Regularised Kernel Subspace Learning* (RKSL) model. The model aims to learn a shared subspace such that when data pairs of the same identities across different camera views are projected into this subspace, they are close to each other, whilst those from distinct people are further-apart. Importantly the model needs to be learned without any cross-view pairwise labels. To achieve this, our model is designed to extract subtle identity-discriminative information from the given unlabelled data via explicitly encoding two types of data affinity constraints into the subspace learning formulation.

Positive soft inter-view correspondence constraint: The unavailability of labelled cross-view pairs motivates us to search for other inter-view information, which is noisy but still useful. Specifically, it is observed that the similarity/affinity measure between two people's images in different views in the visual feature space contains some noisy but identity-discriminative information. This corresponds to a basic assumption that two visually similar people are more likely to be the same person than two visually dissimilar people. This assumption would hold true in most cases. It underpins our *soft cross-view correspondence* constraint which states that the soft cross-view correspondence relationship needs to be preserved in the learned subspace. This constraint is much softer, compared to the labelled hard correspondence constraint exploited by most supervised distance metric learning models.

Negative intra-view affinity constraint: In contrast to the inter-view relationship which we want to preserve in the subspace, we wish two visually similar people (i.e. close in the visual feature space) in the gallery set are separated in the subspace. This constraint is thus to break the

/* *Variables associated with view* $\mathcal{X}$ (similar for $\mathcal{Y}$) */:
$\{\breve{\mathbf{x}}_i\}_{i=1}^{n_u}$: Unlabelled data, with feature matrix $\breve{X}$;
$\{\bar{\mathbf{x}}_i\}_{i=1}^{n_l}$: Labelled data, with feature matrix $\bar{X}$;
$\{\hat{\mathbf{x}}_i\}_{i=1}^{n_x} = \{\breve{\mathbf{x}}_i\}_{i=1}^{n_u} \cup \{\bar{\mathbf{x}}_i\}_{i=1}^{n_l}$, and $\hat{X} = [\bar{X};\breve{X}]$;
$\mathcal{L}_{\hat{x}}$: Graph Laplacian matrix;
$\mathbf{w}_x$: Projection vectors (model parameters);
α: Kernelised projection vectors;
/* *Variables across view* $\mathcal{X}$ *and* $\mathcal{Y}$ */:
S_{ij}: Similarity measure between $\breve{\mathbf{x}}_i$ and $\breve{\mathbf{y}}_j$
/* *Others* */:
K: Kernel matrix on data, further clarified by subscript.

Figure 3.3: Definition of notations.

local affinity structure within each view (see Fig. 3.2 the dashed lines). This constraint is related to the inter-class constraints in classic techniques such as Fisher discriminative analysis, and is designed to make the people on the watch list more distinguishable in the learned subspace. Note, this information is readily available given the one-shot images of different people in a gallery view, and does not require any labelling.

3.3.1 Model Formulation

Formally, with the two constraints described above formulated as two regularisation terms respectively, our RKSL model has the following objective function:

$$\rho = \max_{\mathbf{w}_x,\mathbf{w}_y} \frac{\mathbf{w}_x^\top (\sum_{i,j} \mathcal{S}_{ij} \cdot \breve{\mathbf{x}}_i \breve{\mathbf{y}}_j^\top) \mathbf{w}_y}{\sqrt{\mathbf{w}_x^\top (C_{\hat{x}\hat{x}} + R_{\hat{x}}) \mathbf{w}_x \, \mathbf{w}_y^\top C_{\hat{y}\hat{y}} \mathbf{w}_y}} \tag{3.1}$$

with
$$\begin{aligned} C_{\hat{x}\hat{x}} &= \hat{X}^\top \hat{X} \\ C_{\hat{y}\hat{y}} &= \hat{Y}^\top \hat{Y} \end{aligned} \tag{3.2}$$
$$R_{\hat{x}} = -\frac{\gamma_x}{n_x^2} \hat{X}^\top \mathcal{L}_{\hat{x}} \hat{X} \tag{3.3}$$

where $C_{\hat{x}\hat{x}}$ and $C_{\hat{y}\hat{y}}$ are the covariance matrices among data for the two views, and other notations are explained in Fig. 3.3. In this subspace learning formulation, each data point represented in a visual feature space $\mathcal{F}$ is projected to a subspace $\mathcal{P}$. The projection is realised by two projection matrices $\mathbf{w}_x$ and $\mathbf{w}_y$ for the two views respectively, which are also the model parameters needed to learn. Note that in the OS^{2}Re-Id setting, $\hat{X} = \breve{X}$, and $\hat{Y} = \breve{Y}$ since no cross-view labelled data is available, i.e. $\bar{X} = \bar{Y} = \emptyset$.

In Eq. (3.1), the nominator $\mathcal{B} = \mathbf{w}_x^\top (\sum_{i,j} \mathcal{S}_{ij} \cdot \breve{\mathbf{x}}_i \breve{\mathbf{y}}_j^\top) \mathbf{w}_y$ enforces the positive soft inter-view correspondence constraint, dictating that the *cross-view* similarity/affinity relationship in $\mathcal{F}$ should be preserved in $\mathcal{P}$. More precisely, the similarity between a cross-view unlabelled data pair $\{\check{\mathbf{x}}_i, \check{\mathbf{y}}_j\}$ in $\mathcal{P}$ is constrained to be consistent with their similarity $\mathcal{S}_{ij}$ in $\mathcal{F}$ during the learning process. The value of $\mathcal{S}_{ij}$ can be set by either learning or non-learning based methods as detailed in Sec. 3.5.

On the other hand, $R_{\hat{x}}$ in the denominator of Eq. (3.1) represents the negative intra-view affinity regularisation for constraining $\mathbf{w}_x$, so that in the gallery camera view $\mathcal{X}$, *intra-view* visually similar person pairs are pulled apart in the subspace. Formally, we denote $A_{\hat{x}}$ as a K-Nearest-Neighbour (KNN) similarity graph on $\hat{X}$. By the properties of graph Laplacian [110], we have:

$$\mathbf{w}_x^\top \hat{X}^\top \mathcal{L}_{\hat{x}} \hat{X} \mathbf{w}_x = \frac{1}{2} \sum_{i,j=1}^{n_x} (\mathbf{w}_x^\top \hat{\mathbf{x}}_i - \mathbf{w}_x^\top \hat{\mathbf{x}}_j)^2 A_{\hat{x}}^{ij} \tag{3.4}$$

where $\mathcal{L}_{\hat{x}}$ is the graph Laplacian matrix of $A_{\hat{x}}$. Therefore, $R_{\hat{x}}$ is then computed as the summation over pairwise distances in space $\mathcal{P}$ on visually alike people from gallery view $\mathcal{X}$ (see Eq. (3.3)). By adding its negative regularisation term onto the denominator of Eq. (3.1), we explicitly enforce the adjacent samples in $\mathcal{F}$ to be more separated in $\mathcal{P}$, and in return make the projection $\mathbf{w}_x$ more identity discriminative.

Interestingly, rather than maintaining the locality manifold structures as in models designed for classification [110, 111], our negative regularisation term $R_{\hat{x}}$ on the gallery set is designed to distort them so as to make the projection directions more distinguishable with respect to identities. This is more appropriate for our verification task. However, we do not intend to completely destroy the local manifold structure by over-distortion. We thus impose this negative affinity constraint only on the most visually similar (so confusing) intra-view pairs by using sparse $A_{\hat{x}}$ (i.e. a small K in the KNN graph[1]), whose effect is further controlled by the weight γ_x. Note that a similar negative constraint can be applied to the probe set if this information is available as explained next.

To further extend our model, let us now consider the situation when some labelled cross-view pairs are available (e.g. as assumed in conventional re-id settings). To that end, we introduce a third regularisation term to represent any pairwise labelled information by expanding Eq. (3.1) as follows:

$$\max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\mathbf{w}_x^\top (\sum_k^{n_l} \bar{\mathbf{x}}_k \bar{\mathbf{y}}_k^\top) \mathbf{w}_y + \eta \cdot \mathbf{w}_x^\top (\sum_{i,j} \mathcal{S}_{ij} \cdot \check{\mathbf{x}}_i \check{\mathbf{y}}_j^\top) \mathbf{w}_y}{\sqrt{\mathbf{w}_x^\top (C_{\hat{x}\hat{x}} + R_{\hat{x}}) \mathbf{w}_x \, \mathbf{w}_y^\top (C_{\hat{y}\hat{y}} + R_{\hat{y}}) \mathbf{w}_y}} \tag{3.5}$$

[1] K is set to 15 in this work.

where $\mathbf{w}_x^\top(\sum_k^{n_l}\bar{\mathbf{x}}_k\bar{\mathbf{y}}_k^\top)\mathbf{w}_y$ is the new regularisation term for encoding the *labelled* cross-view data pairs. The coefficient η is a balancing weight parameter for controlling the trade-off between the hard and soft cross-view data correspondences. Note that we also introduce the negative regularisation term $R_{\bar{y}} = -\frac{\gamma_y}{n_{\bar{y}}^2}\bar{Y}^\top\mathcal{L}_{\bar{y}}\bar{Y}$ (similar to $R_{\hat{x}}$ in Eq. (3.3)), for the probe set data whose identities are given from the cross-view pairwise labels.

The objective function in Eq. (3.5) assumes linear projections. However, given significant changes across views in lighting conditions, poses, and occlusions, the optimal subspace for cross-view matching may not be obtainable by linear projections. We thus further kernelise Eq. (3.5) by projecting the data from the original visual feature space into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with an implicit feature mapping function $\phi(\cdot)$. The inner-product of two data points in $\mathcal{H}$ can be computed by a kernel function K, with $K(\mathbf{x}_i,\mathbf{x}_j) = \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$. With the 'kernel trick' [112], we obtain the kernelised objective function as:

$$\max_{\alpha,\beta}\frac{\alpha^\top K_{\hat{x}\bar{x}}K_{\bar{y}\hat{y}}\beta + \eta\cdot\alpha^\top(\sum_{i,j}\mathcal{S}_{ij}\cdot K_{\hat{x}\breve{x}_i}K_{\breve{y}_j\hat{y}})\beta}{\sqrt{\alpha^\top(K_{\hat{x}\hat{x}}^2+\mathcal{R}_{\hat{x}})\alpha\,\beta^\top(K_{\hat{y}\hat{y}}^2+\mathcal{R}_{\bar{y}})\beta}} \tag{3.6}$$

where α and β are the kernelised projection vectors for the two views respectively, and the kernelised $\mathcal{R}_{\hat{x}}$ and $\mathcal{R}_{\bar{y}}$ are:

$$\begin{aligned}\mathcal{R}_{\hat{x}} &= \varepsilon_x K_{\hat{x}\hat{x}} - \frac{\gamma_x}{n_x^2}K_{\hat{x}\hat{x}}\mathcal{L}_{\hat{x}}K_{\hat{x}\hat{x}},\\ \mathcal{R}_{\bar{y}} &= \varepsilon_y K_{\hat{y}\hat{y}} - \frac{\gamma_y}{n_y^2}K_{\hat{y}\bar{y}}\mathcal{L}_{\bar{y}}K_{\bar{y}\hat{y}}.\end{aligned} \tag{3.7}$$

To prevent potential issues caused by the high-dimensional feature maps $\phi(\cdot)$, we introduce $\varepsilon_x K_{\hat{x}\hat{x}}$ and $\varepsilon_y K_{\hat{y}\hat{y}}$ to penalise the norms of the associated projection vectors respectively, which is equivalent to Tikhonov regularization [112]. In our evaluation, we set both ε_x and ε_y to the standard value of 0.5 [112], and utilised the exponential chi-square kernel function.

Now after kernelisation we obtain our final subspace learning model (Eq. (3.6)) termed *Regularised Kernel Subspace Learning* (RKSL). Among the three regularisation terms in Eq. (3.6), (1) the term $\alpha^\top(\sum_{i,j}\mathcal{S}_{ij}\cdot K_{\hat{x}\breve{x}_i}K_{\breve{y}_j\hat{y}})\beta$ utilises unlabelled cross-view data to enforce the *positive soft cross-view correspondence constraint*; (2) the term $\mathcal{R}_{\hat{x}}$ / $\mathcal{R}_{\bar{y}}$ uses intra-view data to enforce the *negative intra-view affinity constraint*, and (3) the term $\alpha^\top K_{\hat{x}\bar{x}}K_{\bar{y}\hat{y}}\beta$ employs labelled cross-view data to enforce the *positive hard cross-view correspondence constraint*. When the cross-view pairwise labels are unavailable, the third term and half of the second term (i.e. $\mathcal{R}_{\bar{y}}$) are removed.

Otherwise, all three terms are kept. This shows the flexibility of our model to deal with different levels of data annotation.

3.3.2 Model Optimisation

We observe that the value of ρ in Eq. (3.6) is not changed when rescaling either α or β or both. Thus, the optimisation problem in Eq. (3.6) is equivalent to maximising its numerator subject to the following two constraints:

$$\begin{aligned} &\alpha^\top (K^2_{\hat{x}\hat{x}} + \mathcal{R}_{\hat{x}})\alpha = 1, \\ &\beta^\top (K^2_{\hat{y}\hat{y}} + \mathcal{R}_{\bar{y}})\beta = 1 \end{aligned} \tag{3.8}$$

So, the corresponding Lagrangian is:

$$\begin{aligned} L(\lambda_x, \lambda_y, \alpha, \beta) = \alpha^T (K_{\hat{x}\bar{x}} K_{\bar{y}\hat{y}} + \eta \cdot \sum_{i,j} \mathcal{S}_{ij} \cdot K_{\hat{x}\breve{x}_i} K_{\breve{y}_j\hat{y}})\beta \\ - \frac{\lambda_x}{2}\big(\alpha^\top (K^2_{\hat{x}\hat{x}} + \mathcal{R}_{\hat{x}})\alpha - 1\big) - \frac{\lambda_y}{2}\big(\beta^\top (K^2_{\hat{y}\hat{y}} + \mathcal{R}_{\bar{y}})\beta - 1\big) \end{aligned} \tag{3.9}$$

By denoting

$$\begin{aligned} &C = K_{\hat{x}\bar{x}} K_{\bar{y}\hat{y}} + \eta \cdot \sum_{i,j} \mathcal{S}_{ij} \cdot K_{\hat{x}\breve{x}_i} K_{\breve{y}_j\hat{y}}, \\ &B_x = K^2_{\hat{x}\hat{x}} + \mathcal{R}_{\hat{x}}, \qquad B_y = K^2_{\hat{y}\hat{y}} + \mathcal{R}_{\bar{y}} \end{aligned} \tag{3.10}$$

where C refers to the *cross-view* term, B_x and B_y the corresponding *intra-view* terms, Eq. (3.9) can be re-written as:

$$L(\lambda_x, \lambda_y, \alpha, \beta) = \alpha^T C \beta - \frac{\lambda_x}{2}(\alpha^\top B_x \alpha - 1) - \frac{\lambda_y}{2}(\beta^\top B_y \beta - 1) \tag{3.11}$$

Setting the derivatives of L in respect to α and β to zeros, we obtain:

$$\frac{\partial L}{\partial \alpha} = C\beta - \lambda_x B_x \alpha \ = 0, \tag{3.12a}$$

$$\frac{\partial L}{\partial \beta} = C^\top \alpha - \lambda_y B_y \beta = 0 \tag{3.12b}$$

By subtracting $\beta^\top$ times Eq. (3.12b) from $\alpha^\top$ times Eq. (3.12a), we have

$$0 = \lambda_y \beta^\top B_y \beta - \lambda_x \alpha^\top B_x \alpha \tag{3.13}$$

After combining Eq. (3.13) with our constraints in Eq. (3.8), we get $\lambda_x = \lambda_y = \lambda$. Together with Eq. (3.12b), we have

$$\beta = \frac{B_y^{-1} C^\top \alpha}{\lambda} \tag{3.14}$$

Substituting Eq. (3.14) into Eq. (3.12a), we get

$$C B_y^{-1} C^\top \alpha = \lambda^2 B_x \alpha \tag{3.15}$$

Thus, we obtain a generalised eigenproblem of the form $A\mathbf{x} = \lambda B\mathbf{x}$. By solving this eigenproblem in Eq. (3.15), we eventually find the closed-form solution of our RKSL model, the optimal projection matrices α and β defined in Eq. (3.6). Specifically, for each eigenvector α and its eigenvalue λ obtained from solving Eq. (3.15), we also get a corresponding paired β with Eq. (3.14).

3.3.3 Model Deployment

Under the OS^{2}Re-Id scenario, given the watch list (gallery set) G and the unlabelled probe set P, we can obtain their representations in the projected space $\mathcal{P}$ by applying the proposed RKSL (Eq. (3.6)) on G and P. This new representation is learned to be identity-sensitive due to the discriminative learning strategies as detailed above. Therefore, we directly use the projected data points in the subspace to perform re-id with the cosine distance [112] as the matching function.

3.4 Datasets and Experimental Settings

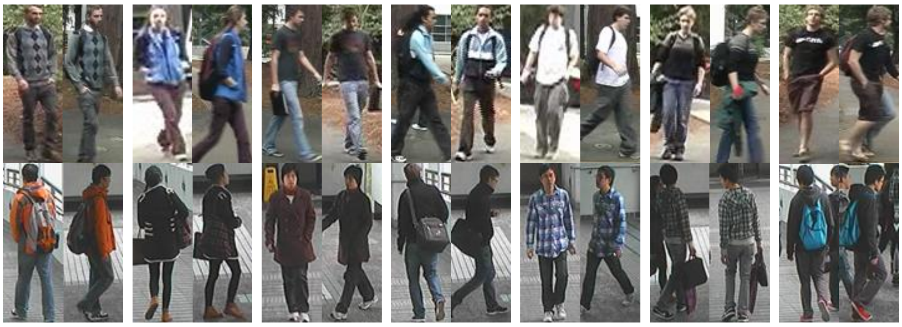


Figure 3.4: Examples of matched cross-view image pairs sampled from the VIPeR [7] (first row) and the CUHK01 [108] (second row) dataset.

Datasets: Under the OS^{2}Re-Id setting, a large probe set is needed for simulating real-world application settings. Therefore, we selected two large benchmark datasets VIPeR [7] and CUHK01 [108], for the evaluation of the proposed RKSL model. Specifically, the VIPeR dataset contains a total of 632 people with one image per person per view, whilst the CUHK01 dataset 971 people with two images per person per view. Both datasets include two disjoint outdoor camera views. These two ReID datasets are challenging due to the large and unknown cross-view variations in view angle, illumination conditions, background clutter and diverse/random occlusion (see Figure 3.4).

Visual features: We adopted the histogram-based image descriptor introduced in [61] as the person appearance representation. Specifically, three types of features were included: (1) Colour histogram: First, the images were segmented horizontally into 15 even overlapped stripes. Second, for each stripe a weighted colour histogram was extracted in each channel of the HS, RGB and Lab colour spaces. Finally, the histogram was then quantised as: 8×8 (HS), $4 \times 4 \times 4$ (RGB), and $4 \times 4 \times 4$ (Lab), resulting in a 2880-D colour vector. (2) HOG [29]: The HOG feature was computed on 8×8 pixel blocks with cell size of 2×2. For each cell the gradients were quantised into 4 bins. (3) LBP [113]: The LBP histogram was calculated on grids sized 16×16. The bin size for quantisation was set to 58. The final image feature vector (5138-D) was obtained as the concatenation of these three histograms.

3.5 Experiments and Evaluations

3.5.1 Unsupervised Re-Identification Evaluation

We evaluated the re-id performance of unsupervised methods under the OS^{2}Re-Id setting where no cross-view labelled data pair is available.

Settings: For both datasets, we created the watch list of target people (gallery set) G by randomly selecting 120 different people from one camera view, and the probe set P by selecting half of the whole population (316 on VIPeR and 486 on CUHK01) from the other view, with the condition that 100 people exist in both G and P. Therefore, there are 216 (= 316-100) imposters in P for VIPeR, and 386 (= 486-100) for CUHK01. For either G or P, only *one-shot image per person* is included[2]. We evaluated a total of 10 folds and reported their averaged results.

Competitors: We compared with four baseline methods: (1) L1-norm[3]: a basic distance metric. (2) SDALF [38]: a type of hand-crafted visual feature specially designed for re-id. (3) SDC [33]:

[2] Even though two shots per person per view are available on CUHK01, we randomly selected and used one of the two.

[3] We found that L2-norm distance gave almost identical results.

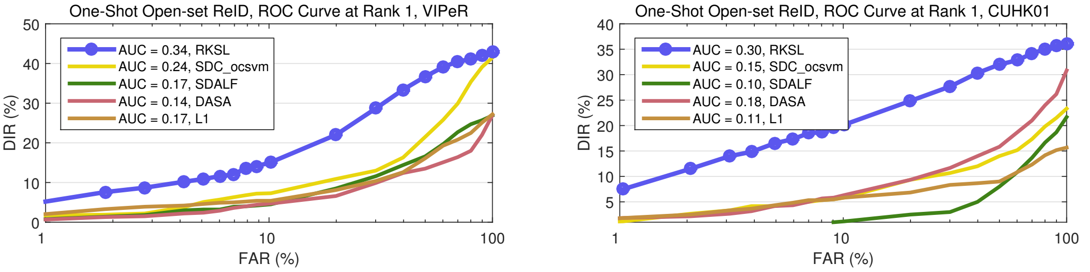


Figure 3.5: Comparing Rank-1 scores of all methods over all FARs on VIPeR (left) and CUHK01 (right). OneShot-OpenSet Re-Id setting.

a state-of-the-art unsupervised re-id model. Note that for mining localised saliency statistics, this model requires two additional reference sets, one for each camera view and containing the same group of (100) people [33]. (4) DASA [109]: a state-of-the-art unsupervised domain adaptation model. In the re-id context, each person is considered as a class, and each camera view as a domain.

Evaluation metric: We utilised the ROC curves of False Accept Rate (FAR) versus Detection and Identification Rate (DIR) for performance comparison [114]. Specifically, two steps are involved: (1) Detection - decide whether a probe person i exists in the gallery based on its estimated similarity measure $\{s_{i,j}\}_{j=1}^{|G|}$ with the gallery and a decision threshold τ, i.e. yes if $\max(\{s_{i,j}\}_{j=1}^{|G|}) > \tau$, no otherwise. (2) Identification - compute the cumulated matching rank rates over accepted target people. Note that DIR becomes the Cumulated Matching Characteristics used for the conventional closed-world setting, when FAR $= 100\%$.

Implementation details: For the parameter setting of our RKSL model, since no labelled data is available, we cannot use cross-validation to tune the model parameters and they have to be set empirically. The only parameter we need to set is $\gamma_x/n_x^2 = 0.02$ (see Eq. (3.3)). The value of K in the KNN graph and ε were set to standard values (15 and 0.5 respectively as in [110, 111]). We found that the result was very insensitive to its value. For computing the soft cross-view correspondence S_{ij}, we simply used the additive inverse of L2 distance between each pair of cross-view unlabelled data, and normalise its value to a range between 0 and 1.

Comparative Results

It is evident from Figure 3.5 and Table 3.1 that the proposed RKSL model significantly outperforms all the competitors on both datasets, particularly with demanding (small) FARs. Particularly, when compared to the second best method (SDC on VIPeR and DASA on CUHK01) at FAR = 10%, the Rank-1 score is doubled (from 7.3 to 15.1) on VIPeR and tripled (from 6.3 to

Dataset	VIPeR				CUHK01			
FAR (%)	1	10	50	100	1	10	50	100
L1-norm	1.9	5.4	16.1	27.2	1.8	5.8	9.0	15.7
DASA[109]	0.6	4.7	13.5	27.0	1.8	6.3	15.8	30.8
SDALF[38]	0.7	4.5	16.6	26.9	0.2	1.2	8.0	21.7
SDC[33]	1.7	7.3	21.5	41.5	1.2	5.8	14.0	23.3
RKSL	**4.9**	**15.1**	**36.7**	**42.9**	**7.5**	**20.2**	**32.0**	**36.0**

Table 3.1: Comparing Rank-1 scores of different methods at varying FARs. OneShot-OpenSet Re-Id setting.

20.2) on CUHK01 by RKSL. This demonstrates the effectiveness of the proposed kernel subspace learning model in extracting identity-sensitive information from the unlabelled data.

We now examine the performance of each individual baseline method. The state-of-the-art unsupervised re-id model, SDC, is shown to be less effective on CUHK01 (with a larger probe set) than VIPeR. A possible explanation is that the 100-people reference sets are not sufficient to capture the localised appearance saliency and more data are needed when a larger population (probe people) is considered. In contrast, the proposed RKSL model can overcome this problem by automatically learning person-discriminative subspace from only unlabelled data using the two complementary pairing constraints (Section 3.3), without any extra manual cost such as the need of constructing a reference set for each view with the same group of people. Compared with RKSL, DASA is much inferior in matching people across views. This suggests that it is difficult for the unsupervised domain adaption approach to solve the re-id problem where the intrinsic discriminative information can be more subtle and more challenging to extract than in the general object recognition/categorisation problem, especially when their assumption on the two domains containing the same set of classes becomes invalid under our setting. Interestingly, on CUHK01 dataset we found that SDALF generates even poorer results than L1 except when FAR = 100% which corresponds to the closed-world setting. This demonstrates the significant challenges of manually designing re-id features, particularly under the more realistic OS^{2}Re-Id setting.

Computational Cost Analysis In addition to re-id accuracy, we also quantitatively compared these methods in terms of efficiency, since it is another important metric to evaluate the usefulness of a model in real-world large scale person re-id application. The running time was measured on a desktop machine with Intel CPU at 3.30 GHz and memory of 8.0 GB with MATLAB implementation for all compared models. This comparison was made on VIPeR. On average, for each fold of re-id experiment, the RKSL model took 0.08 minute (4.8 seconds), whilst SDC 104.26 minutes and SDALF 173.18 minutes. In other words, RKSL is >1000 and >2000 times faster than SDC and SDALF respectively. This confirms the greater suitability of the proposed RKSL model over its competitors for the large scale and real-time re-id application in reality.

3.5.2 Semi-Supervised Re-Identification Evaluation

In addition to OS^{2}Re-Id, we also wish to investigate the effects of accommodating labelled data in model learning. We thus extensively compared the effectiveness of the proposed RKSL model with existing re-id methods in the conventional semi-supervised settings where some (sparse) cross-view labelled pairs are available.

Settings: We followed the same semi-supervised setting as in [70]. Specifically, for either VIPeR or CUHK01, we split the whole dataset into two partitions: one half for training and the other half for testing. One third of the training partition are cross-view pairwise labelled. For a fair comparison, on CUHK01 dataset the multi-shot matching as in [115, 21] was adopted for all comparative methods in this semi-supervised setting.

Competitors: We compared the RKSL model with the only comparable semi-supervised re-id method, SSCDL [70], as well as four most contemporary fully-supervised models including, RankSVM [36], KISSME [18], kLFDA [23], and KCCA [61]. For a fair comparison, we utilised the same visual feature in all methods, except SSCDL which is a patch-based matching approach and thus their reported results were compared.

Evaluation metric: The conventional Cumulated Matching Characteristics (CMC) curves were utilised for quantitative comparison between different methods.

Implementation details: Under this semi-supervised setting, we used cross-validation to determine the free parameters $(\eta, \gamma_x, \gamma_y)$ for the proposed RKSL model, as well as parameters of all the baseline methods [36, 18, 23, 61].

Comparative Results

The results of all compared methods on both datasets are shown in Figure 3.6 and Table 3.2. It is observed that the proposed RKSL model significantly outperforms all baseline methods, particulalry at the top ranks. Specifically, RKSL provides much better re-id accuracy than the state-of-the-art semi-supervised model SSCDL, e.g. a $\sim 9\%$ absolute improvement at Rank-1. In general, performance gains on top ranks are regarded more important and desirable in practical

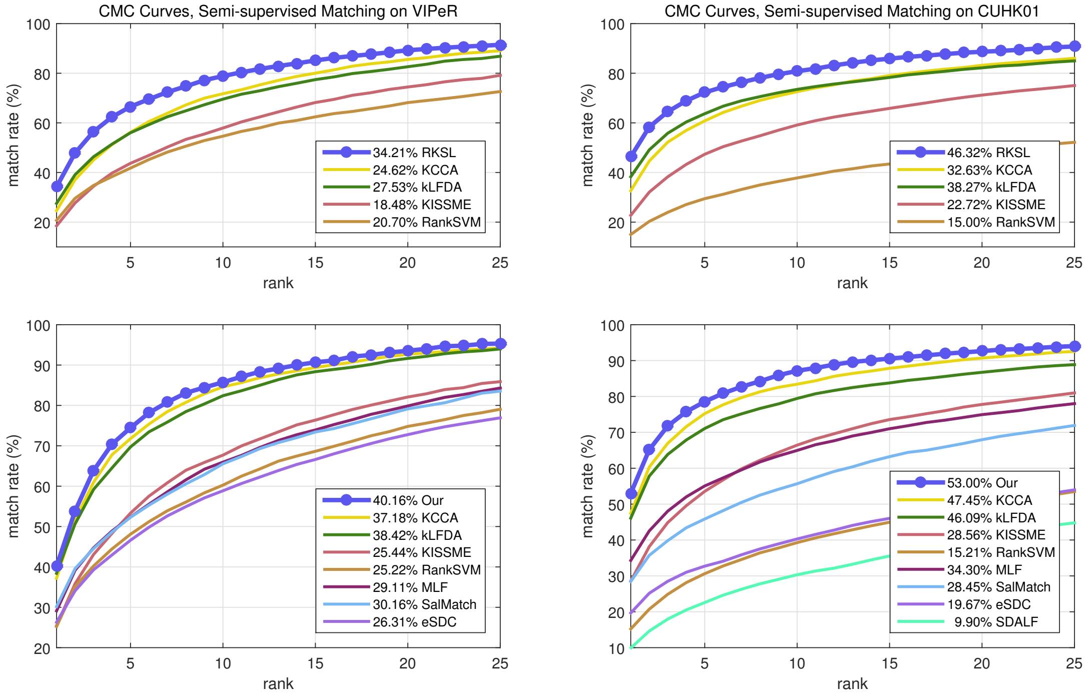


Figure 3.6: Comparing the performance of different methods on VIPeR (left) and CUHK01 (right). The semi-supervised re-id setting (top row). We also include here our evaluation results under the conventional fully-supervised setting (bottom row).

ReID applications, particularly for Rank-1. This shows the effectiveness of our RSKL method in learning identity-discriminative information from both unlabelled and labelled data using a unified single formulation integrating three types of pairwise relationships simultaneously (see Figure 3.2 and Eq. (3.6)). In particular, this demonstrates the importance of our soft cross-view correspondence constraint over unlabelled data for cross-view people matching, which however is totally ignored by SSCDL for its model learning/optimisation.

The results also show that all fully-supervised models yield much worse recognition results than RKSL. For example, for VIPeR our RKSL improves Rank-1 score over RankSVM [36] by

Dataset	VIPeR				CUHK01			
Ranks	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
RankSVM[36]	20.70	41.77	54.62	68.16	15.00	29.44	37.79	48.18
KISSME[18]	18.48	43.70	57.90	74.46	22.72	47.37	59.13	71.19
kLFDA[23]	27.53	56.01	69.55	82.62	38.27	63.68	73.49	82.18
KCCA[61]	24.62	56.20	71.74	85.56	32.63	60.80	72.57	83.21
SSCDL[70]	25.60	53.70	68.10	83.60	-	-	-	-
RKSL (Ours)	**34.21**	**66.55**	**78.86**	**89.27**	**46.32**	**72.28**	**80.82**	**88.66**

Table 3.2: Comparing some matching rates of different methods on VIPeR and CUHK01. The semi-supervised re-id setting.

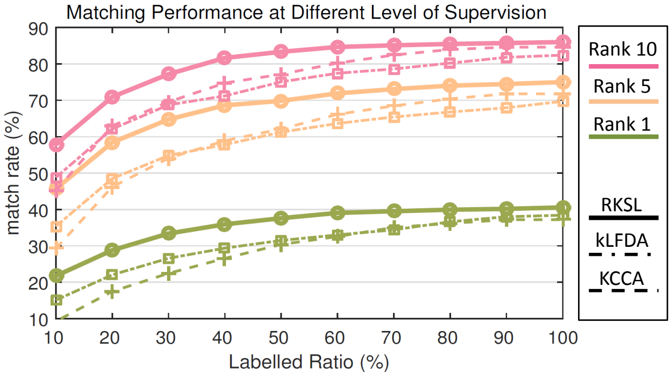


Figure 3.7: Semi-supervised matching on VIPeR. Matching rate as a function of labelled data percentage.

13.51%, KISSME [18] by 15.73%, kLFDA [23] by 6.68%, and KCCA [61] by 9.59%. And even larger Rank-1 improvements are gained by RKSL on CUHK01. The main reason of inferior performance by these supervised methods is the limited availability of labelled data and their inability of exploiting the large quantity of unlabelled data. Whilst the proposed RKSL model can effectively utilise both in a unified way, largely relaxing the stringent assumption on labelled data amount and making it flexible in coping with varying amounts of data annotation.

Effect of Labelled Data Sparsity

For evaluating the performance given different amount of data annotation, we further conducted a set of experiments on VIPeR by comparing RKSL with the two best baselines, kLFDA [23] and KCCA [61], when different numbers of labelled pairs are provided. To this end, we changed the labelled data percentage from 10% to 100% and compared their performances on several ranks (Rank-1, 5, 10). The results in Figure 3.7 show that the accuracies achieved by the proposed RKSL model are significantly better at all three ranks, compared to the two baselines. The margins are evidently larger when fewer labelled data are available, which further suggests the effectiveness of our RSKL in exploiting unlabelled data for person-discriminative subspace learning. Note that at 100%, this becomes the standard fully supervised re-id setting. Our model operates under this setting by setting $\eta = 0$, i.e. removing the soft cross-view correspondence constraint as no unlabelled data is available. Figure 3.7 shows that our model, although being unable to exploit the unlabelled data now, still outperforms the state-of-the-arts (see Figure 3.6 at Rank-1, RSKL:40.16%, KLFDA:38.41%, and KCCA:37.18%). This further demonstrates the strength and flexibility of our model under a large spectrum of settings.

3.6 Summary

In this Chapter we have presented an unsupervised and open-world re-id setting termed as *OneShot-OpenSet Re-Id* (OS^{2}Re-Id). To solve the problem, a novel *Regularised Kernel Subspace Learning* (RKSL) model is proposed. The model is unique due to its capability of learning cross-view identity-discriminative information from unlabelled data. This characteristics makes RKSL readily applicable and scalable to large scale re-id problems. Also, the RKSL model allows to effectively exploit pairwise labels when available. Extensive comparative evaluations were conducted to validate the advantages of the proposed model in both under the OS^{2}Re-Id (no pairwise labels) and conventional (with labelled data) settings.

Chapter 4

Saliency Discovery from Unlabelled Data

4.1 Overview

In this chapter we continue to investigate the question that what can be explored from unlabelled data for model training in order to save human labelling efforts. While Chapter 3 has proposed one possible solution to the problem by learning a global matching function from unlabelled data, this Chapter explores a different strategy by looking into the localised regions of unlabelled person images. In particular, this chapter proposes a novel unsupervised re-id modelling approach by exploring generative probabilistic topic modelling. Given abundant unlabelled data, our topic model learns to simultaneously both (1) discover localised person foreground appearance saliency (salient image patches) that are more informative for re-id matching, and (2) remove busy background clutters surrounding a person. Extensive experiments are carried out to demonstrate that the proposed model outperforms existing unsupervised learning re-id methods with significantly simplified model complexity. In the meantime, it still retains comparable re-id accuracy when compared to the state-of-the-art supervised re-id methods but without any need for pair-wise labelled training data.

4.2 Problem Definition

Recent efforts on solving the re-id problem are dominated by supervised learning based methods that aim to learn an optimal matching function or distance metric [19, 34, 36, 53, 116]. More specifically, for each pair of camera views, a *labelled* training set is constructed. It consists of

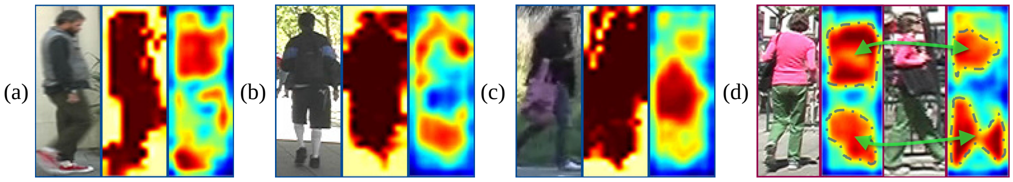


Figure 4.1: Each of (a)-(c) shows (left to right): person image, topic model detected background map and foreground saliency map. The saliency maps capture localised appearance features (e.g. brown jacket, red shoes, blue sleeve pattern, pink handbag, green bottom, pink shirt). (d) show that the distributions of the foreground saliency maps from two different camera views of the same person are stable and useful for re-id. Best viewed in colour.

a set of people for which images of each individual must be annotated manually with an identity label across both views. A matching function is learned from the training set subject to a set of constraints, that is, a pair of images of the same person should have larger matching score/smaller similarity distance compared to that of two different people given the labelling information, regardless their visual appearance dissimilarity/similarity. By satisfying these constraints the learned model can implicitly discover visual features that are more stable against intra-class appearance variations. These variations are typically caused by viewing condition changes across a particular pair of camera views. However, there is a significant limitation of these supervised learning based methods – a large set of people must be labelled manually across every pair of camera views. Moreover, even for the same pair of camera views, once the conditions change (e.g. different time of the day), new labelling may be needed again to update the matching function. Therefore, such approaches are inherently limited in their scalability to different camera pairs at different times without the need for exhaustive and repeated manual labelling. This is impractical for large camera networks of hundreds of cameras.

Based on the reasons stated above, unsupervised methods are thus more preferred for overcoming the limitations of supervised learning. As already been discussed in Section 2.2.3, saliency-based feature selection has been proved to have good properties which can be explored through unsupervised learning, whereas existing saliency-based re-id methods are still imperfect to a large extent. In this chapter, a novel unsupervised modelling approach to saliency detection for person re-id is proposed based on probabilistic generative topic modelling. This is significantly different from previous attempts, which are data-driven and discriminative. More specifically, given abundant unlabelled data, our model aims to learn simultaneously what people look like (background removal in a bounding box) and how their typical appearance can be represented by a collection of local and visually coherent parts. This is achieved by learning a set of latent topics that correspond to both typical and localised human appearance components, e.g. blue jeans and dark suit. This component-based typical appearance representation is then deployed for identifying *atypical* appearance by discovering local saliency. This generative topic model based representation is also inherently capable of differentiating background clutters from typical human appearance in a detected person bounding box (Figure 4.1), beneficial to person re-id in cluttered scenes [38, 37].

This chapter proposes a Generative Topic Saliency (GTS) model based on unsupervised topic modelling designed specifically to discover re-id relevant saliency that corresponds to atypical appearance of individual people (foreground). It also simultaneously removes surrounding background clutter in a person detection bounding box. It has two advantages over the existing saliency model for person re-id [33]: (1) Interpretability - each learned topic has clear semantic meaning. (2) Complexity - only a single model is needed for computing saliency for all the images in a camera view, in contrast to having to construct a different saliency model for every image patch of every image. Comparative evaluations on the VIPeR [7] and iLIDS [8] datasets demonstrate that the proposed GTS model not only outperforms existing unsupervised learning based saliency model, but also is competitive to the state-of-the-art supervised learning models without the need for expensive data labelling.

Contributions – Our contributions are: (1) A novel re-id model, Generative Topic Saliency (GTS), for localised human appearance saliency selection by exploiting unsupervised generative topic modelling. (2) The GTS model is capable of simultaneous foreground saliency detection and background clutter removal. (3) The GTS model yeilds state-of-the-art re-id performance against existing unsupervised learning based re-id methods.

4.3 Unsupervised Saliency Discovery by Generative Topic Modelling

4.3.1 Image Representation

Similar to [33], we adopt an over-sampled local patch based representation for each person image. More precisely, each image is represented by 50% overlapped uniform-sized square patches on a dense grid. From each patch, a 32-bin color histogram is computed in the LAB color space with 3 levels down-sampled. SIFT features are also computed in the 3 color channels, with each patch divided into 4×4 cells and 8-bin orientations of local gradients. The final patch descriptor is computed by $L2$ normalisation and concatenation of the colour histogram and SIFT, giving a

672 dimensional feature vector ($32\times3\times3+8\times4\times4\times3$). Patch size and grid step length are 10 and 4 pixels respectively. Our overall image representation builds on the patch descriptors and differs from that of [33]. Specifically, a topic model treats each document (image) as a certain combination of visual words and requires a bag-of-words representation. Given the patch feature vectors from each image, we cluster all the patch feature vectors from an unlabelled training set into a $N_v = 2000$ words codebook by K-means clustering. Given this codebook, each patch is assigned with a word label by its cluster index. An image I_m is then represented by N_m words together with their image positions, denoted as $\{w_{nm}, l_{x_{nm}}, l_{y_{nm}}\}_{n=1}^{N_m}$, with w_{nm} the word label of a patch, l_{x_ij} and $l_{y_{ij}}$ the image coordinates of that patch.

4.3.2 Model Formulation

Given a set of M images of people in bounding boxes, typically extracted from a person detector, we wish to learn a joint topic model capable of capturing the typical appearance of people in foreground patches and simultaneously separating the background patches within each bounding box, without any labelling information. The topic model essentially factorises the image patches and attempts to find localised coherent patches (not necessarily connected) that correspond to common appearance traits of people such as grey top and blue jeans, without any supervised learning. However, the bounding boxes inevitably contain backgrounds, which are often also spatially and visually coherent. To differentiate them, background patches are also modelled explicitly by the generative topic modelling. We thus learn two types of latent topics in our model corresponding to foreground and background respectively. Since foreground appearance are in general more 'compact' than background, similar to [117] we choose a Gaussian distribution to encode foreground human appearance topics and a Uniform distribution to encode more spread-out background topics.

Model Description – Our model is a generalisation of the Latent Dirichlet Allocation (LDA) model [118] with an added spatial variable to make the learned topics spatially coherent. Given a dataset of M images, each image will be factorised (clustered) into a unique combination of K shared topics, with each topic generating its own proportion of words on that image. Conceptually, one topic encodes a certain distribution of visual words (patches), whose vocabulary and spatial location revealing certain patterns, in our case the visual characteristics of human appearances and backgrounds. Among these K topics, K^{ha} topics are used to model foreground human appearance, and $K^{cb} = K - K^{ha}$ topics represent background within the bounding boxes from the entire training dataset. In this work we set $K^{cb} = K^{ha} = 20$ as in [117]. Suppose *Dir*, *Multi*, $\mathcal{NW}$, $\mathcal{N}$ denote respectively Dirichlet, Multinomial, Normal-Wishart and Normal distributions, the generative process of this model is:

1. For each topic $t_k \in \{t_1, t_2, \dots, t_K\}$, draw its appearance distribution $\beta_k \sim Dir(\beta_k^0)$.
2. For each image $I_m \in \{I_1, I_2, \dots, I_M\}$, draw the human appearance and camera background topic distribution $\theta_m \sim Dir(\alpha)$. Each human appearance topic $t_k \in T^{ha}$ is assigned with a Gaussian distribution parameters to reflect the spatial location and size of the human appearance on I_m: $\{\mu_{km}, \sigma_{kj}\} \sim \mathcal{NW}(\mu_0^k, \lambda_0^k, W_0^k, v_0^k)$.
3. For each patch $P_{nm} \in \{P_{1m}, P_{2m}, \dots, P_{N_m m}\}$, draw its topic $z_{nm} \sim Multi(\theta_m)$, draw its vocabulary $w_{nm} \sim Multi(\beta_{z_{nm}})$ and draw its location $\mathbf{l}_{nm}$. If z_{nm} is a human appearance topic then its location is Gaussian distributed, $\mathbf{l}_{nm} \sim \mathcal{N}(\mu_{z_{nm}m}, \sigma_{z_{nm}m}^{-1})$; if z_{nm} is a camera background topic then its location is Uniformly distributed, $\mathbf{l}_{nm} \sim Uniform$.

Model Learning – The learning task for this model is to infer the following quantities: (1) The vocabulary distribution of each human appearance and background topics β_k, (2) all topics' word proportion θ_m and their locations $\{\mu_{mk}, \sigma_{mk}\}$ in each image, and (3) each patch's topic assignment z_{nm}. The joint distribution of observed data set O, latent variables set L and hyper-parameters set H is given by:

$$\begin{aligned} Pr(O,L|H) = \prod_m^M \prod_k^K \Bigg[& Pr(\mu_{mk}, \sigma_{mk} | \mu_0^k, \lambda_0^k, W_0^k, v_0^k) Pr(\theta_m | \alpha_m) \\ & \left(\prod_n^{N_m} Pr(w_{nm} | z_{nm}, \theta_m) Pr(z_{nm} | \theta_m) \right) \Bigg] Pr(\beta_k | \beta_k^0) \end{aligned} \tag{4.1}$$

This model is intractable by exact solutions. An approximate solution can be learned by the EM algorithm with a variational inference strategy, through introducing a Dirichlet parameter γ and a multinomial parameter φ as variational parameters. Under this variational inference framework, γ is learned for each image, with γ_{mk} modelling the proportion of patches which belong to topic t_k in image I_m. φ is learned for each patch, with φ_{nmk} modelling the probability of patch P_{nm} on image I_m being generated by topic t_k. The hyper-parameter α is set to 1 for all human appearance and camera background topics because our method is completely unsupervised and thus all topics may appear in any images.

4.3.3 Saliency Discovery

A key objective of our model is to discover local foreground patches in a person's image that make the person stand out from other people, i.e. the model seeks not only visually distinctive but also *atypical* localised appearance characteristics of a person. To compute such a saliency value, let us first consider to compute a 'prevalence' value of each patch and define saliency as the inverse of prevalence, as the former is more naturally computable by the topic model. Specifically, for a patch P_A on image I_A, its saliency value is measured by how *unlikely* this patch will appear in a training set $\mathcal{I}^R$ of J images at the proximity of a particular spatial location in the images. P_A's saliency score is the inverse of its prevalence value in $\mathcal{I}^R$. For computing patch prevalence value, suppose the learned latent variables set is L and their hyper-parameter set is H. The topic appearance vector β_{vk} reflects the probability that vocabulary (the collection of words in the codebook) v is generated under topic t_k. The multinomial parameter φ_{nmk} refers to the probability that patch P_{nm}'s topic is t_k given the learned model parameters:

$$\beta_{kv} = Pr(w = v|t_k, L, H),\ v = 1,2,\ldots,N_v; \quad \varphi_{nmk} = Pr(z_{nm} = t_k|L, H),\ k = 1,2,\ldots,K \tag{4.2}$$

Based on the Bayes' Theorem, combining the two equations in Eqn. (4.2) gives the joint likelihood of observed word w_{nm} and its topic z_{nm} as:

$$Pr(w_{nm} = v, z_{nm} = t_k|L, H) = Pr(w = v|t_k, L, H)Pr(z_{nm} = t_k|L, H) \tag{4.3}$$

By margining out the topic variable z_{nm} over t_1 to t_K, we obtain the likelihood of patch P_{nm}'s vocabulary w_{nm}. This likelihood value reflects our model's confidence for the visual word w_{nm} to be vocabulary v: $(v = 1,2,\ldots,N_v)$:

$$\mathcal{L}(w_{nm}) = Pr(w_{nm} = v|L, H) = \sum_{z_{nm}=t_1}^{t_K} Pr(w_{nm} = v, z_{nm} = t_k|L, H) \tag{4.4}$$

To measure the probability of patch P_A appearing in image I_m, we impose a simple but reasonable human prior knowledge on people's images, that is, a person's position within a bounding box is relatively stable, and a patch's horizontal shift caused by viewpoint change is far larger than its vertical shift. This assumption is typically valid for a pedestrian captured in a bounding box. Based on this assumption, in each image I_m in $\mathcal{I}^R$ we build a patch set $\hat{P_m^A}$ by taking all the patches in the same horizontal row as P_A. The elements in $\hat{P_m^A}$ are referred as $P_{m,r}^A$, with r as the row index. Given P_A's vocabulary $w_{P_A} = v_0$, the probability that patch P_A repeatedly appears in image I_m of $\mathcal{I}^R$ is measured by the maximum probability for $\hat{P_m^A}$ patches' vocabulary equalling to v_0:

$$\mathcal{P}(P_A \text{ in } I_m) = \max\Big(Pr\Big(w_{P_{m,r}^A} = v_0|L,H\Big)\Big), \quad P_{m,r}^A \in \hat{P_m^A} \tag{4.5}$$

Patch P_A's prevalence level is computed by accumulating $\mathcal{P}(P_A \text{ in } I_m)$ for all the images I_m in $\mathcal{I}^R$:

$$Prevalence(P_A) = \sum_{I_m} \mathcal{P}(P_A \text{ in } I_m), \quad I_m \in \mathcal{I}^R \tag{4.6}$$

Given the prevalence value of each patch (Eqn. (4.6)), its saliency score is initialised by applying an inverse function $h(x)$ on its prevalence value. These saliency scores are then further refined by two basic principles as follows. First, a patch with high probability of belonging to background topics should have low saliency scores. Second, even if a patch belongs to a human appearance topic, but if this topic is very dominant/popular in the training dataset (e.g. many people wearing jeans), the patch also should have low saliency score.

The learned Dirichlet parameter γ_{mk} reveals the proportion of patches on I_m belonging to topic t_k, which can be treated as a pseudo count for the amount of patches falling into each topic on I_m. We then model the popularity of topic t_k by accumulating γ_{mk} over all images in the probe set $\mathcal{I}^p$ and gallery set $\mathcal{I}^g$:

$$Popularity(t_k) = \sum_{I_m} \gamma_{mk}, \quad I_m \in \{\mathcal{I}^p, \mathcal{I}^g\}, t_k \in T^{ha} \tag{4.7}$$

The M foreground topics with highest $Popularity$ values is treated as popular human appearance topics, and deployed to form a topic set T^{pop}. In practice, we take $M = K^{ha}/2$, i.e. 50% of all human appearance topics with higher popularity scores are considered to be statistically common/typical. The final saliency score of patch P_A is computed by combining its prevalence level, the probability of its topic *not* belonging to a background topic, and being less popular (atypical) among foreground appearance topics, i.e.

$$\begin{aligned} Saliency(P_A) = h(Prevalence(P_A)) - \eta_1 \cdot \sum_{t_k \in T^{cb}} Pr(z_A = t_k|L,H) \\ - \eta_2 \cdot \sum_{t_k \in T^{pop}} Pr(z_A = t_k|L,H), \qquad 0 < \eta_1, \eta_2 < 1 \end{aligned} \tag{4.8}$$

where $h(x)$ is a inverse function defined as taking the additive inverse and normalising the result into the $[0,1]$ interval. $Prevalence(P_A)$ is given by Eqn. (4.6). The last two terms can be calculated through Eqn. (4.2), whilst η_1, η_2 are their weights to affect the saliency score, determined by cross-validation during our experiment. If one considers that $Prevalence(P_A)$ simply measures how likely the exact same patch appears repeatedly across images, its topic's *popularity* takes much larger amounts of patches into consideration. These patches may even be visually different from P_A, but they are inherently related by the same topic. This model avoids the topic being simply data-driven; it also considers more inherent structure of the large-scaled data. It is worth pointing out that the model of [33] selects two independent reference training datasets (one for the gallery camera view and another for the probe camera view) and trains many patch-specific and view-specific discriminative models: a different model for every patch of every probe image and every gallery image in order to match the probe image against a set of gallery images for re-id. In contrast, our method only requires to train a *single* model for each camera view given an independent training dataset from that view. Then only two models are required for all patches of all the probe images and all the gallery images respectively. Some examples of the saliency maps obtained using our method are shown in Figure 4.2 and Figure 4.3. In addition, we also show in Figure 4.4 different background patterns discovered by our GTS model, which can be removed for better re-id matching performance.

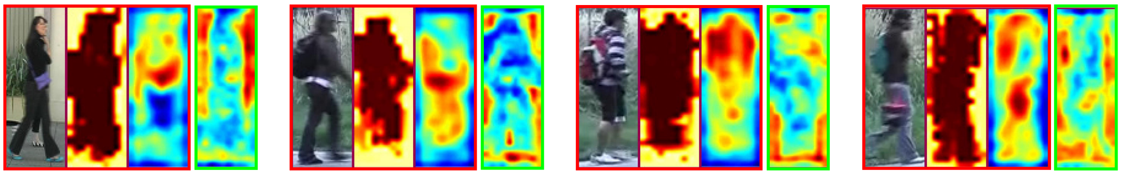


Figure 4.2: Saliency maps comparison (left to right): A person image in detected bounding box, GTS-computed background map, GTS-computed saliency map, saliency map computed by the model of [33] (green bounding box).

4.3.4 Model Deployment

Given the saliency score, we adopt the same patch-based image matching scheme of [33] to compute a matching score between a set of gallery images and a probe image from an independent test set. First we build a corresponding pairwise relationship for all the patches in a probe image I_A and a gallery image I_B. In each patch pair (image location indexed), one patch P_1 is from I_A and the other P_2 from I_B. More precisely, a pair of (P_1, P_2) patch is the nearest neighbour match searched in the proximity of P_1 in I_B or vice versa P_2 in I_A. The matching similarity distance metric is given by $s = exp(-d^2/2a^2)$, where d is the Euclidean distance between two patch feature vectors and a is the bandwidth of a Gaussian function. The overall similarity between the two images is computed by a weighted sum accumulating all the patch pairs' similarities weighted by the saliency scores of patches in each pair, i.e. an accumulation over the quantity $Saliency(P_1) \cdot s(P_1, P_2) \cdot Saliency(P_2)$, where P_1 and P_2 are two patches in one pair. It is worth pointing out that the published code of [33] utilizes foreground masks to remove background patches in VIPeR images. The similarity score between a pair of images is only computed in the foreground region. A similar process of background removal is adopted by many existing works [38, 37, 33]. Body parts information are not explored in our experiments.

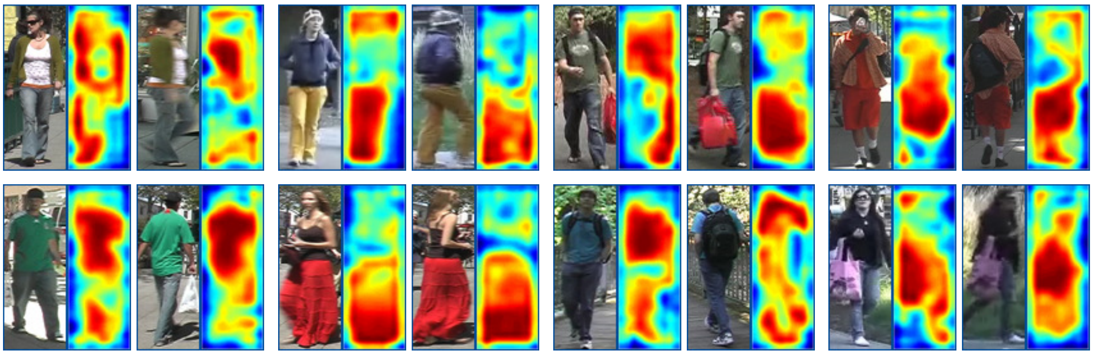


Figure 4.3: More qualitative results of the discovered saliency regions by our unsupervised GTS model. Cross-view image pairs of the same identities with their saliency maps are shown here. The detected saliency regions are found to be stable under different camera views.

4.4 Datasets and Experimental Settings

We evaluate our method on two widely used benchmark datasets, VIPeR [7] and iLIDS [8]. The VIPeR dataset contains 632 pedestrian image pairs. Each pair of images contain the same individual, but were taken from different camera views. Following the experimental setting of [35, 38], we randomly choose half of the dataset, i.e. 316 image pairs, as our training sets. On this training set, we train two topic models, one for each camera view. Among the 316 pairs of training images, we choose 100 pairs as our reference sets for computing saliency and use one reference set per camera view, same as [33]. The iLIDS dataset contains 476 images of 119 people. We followed the same single shot experiment protocol as [53], i.e. randomly choose all images of $p = 50$ people as test set, and use the other images as training set. In the test set, one image per person is chosen to form a gallery set, while all the remaining images compose a probe set. We

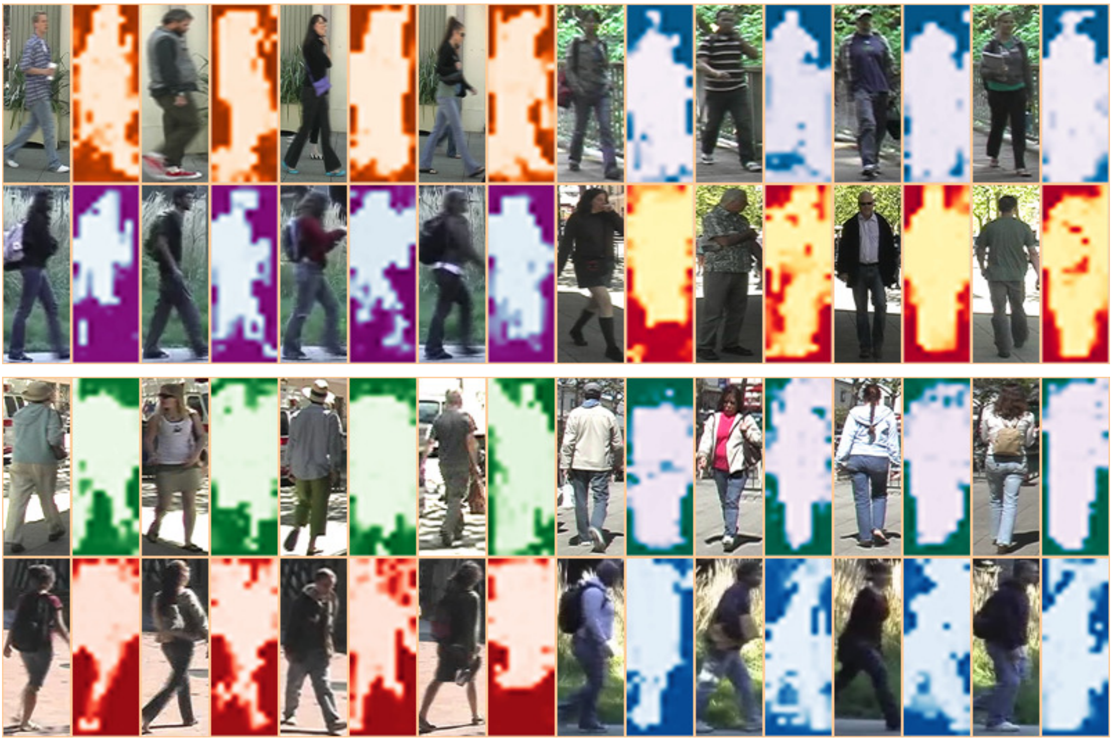


Figure 4.4: Some typical background patterns discovered by GTS, different colors showing different background topics. Coloured regions shows high probability to belong to the topic.

run our experiments for 10 trials with different splits, and report the average of these 10 trials as our final result. The performance is evaluated using the Cumulated Matching Characteristics (CMC) curves.

4.5 Experiments and Evaluations

4.5.1 Unsupervised Competitors Evaluation

We first compare our GTS model against non-learning based methods, i.e. template matching with a distance measure. L1-norm and L2-norm distances are used as the baseline models for comparison. Figures 4.5 and 4.6 show respectively the results on VIPeR and iLIDS. It is evident that our method significantly outperforms the baseline non-learning methods, e.g. Rank-1 about 150% (VIPeR) and 14% (iLIDS) relative improvement over L1-norm. This suggests that the unlabelled data indeed helps improve re-id matching accuracy.

Next we compare GTS to a number of contemporary unsupervised learning methods including eSDC_knn [33], eSDC_ocsvm [33] [1], LDFV [44] and SDALF [38]. Figures 4.5 and 4.6

[1] The results of KNN and OCSVM in our experiments are obtained by running the author published code under our experiment settings. The results are thus slightly different from those reported in [33].

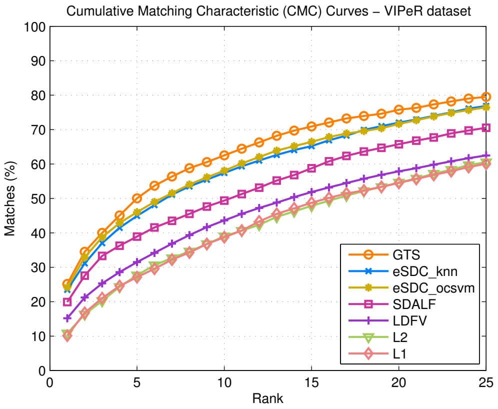


Figure 4.5: VIPeR test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
ELF	12.00	31.50	44.00	61.00
PRDC	15.66	38.42	53.86	70.09
PCCA	19.27	48.89	64.91	80.28
LMNN-R	20.00	49.00	66.00	79.00
KISSME	19.46	48.10	62.50	78.32
RPLM	27.00	-	69.00	83.00
LF	24.18	-	67.12	-
GTS	25.15	50.03	62.50	75.76

Table 4.1: VIPeR test: Comparing the GTS model to supervised learning based models.

show that our model is clearly superior to LDFV and SDALF, e.g. Rank-1 27% (VIPeR) relative improvement over SDALF. These results show that modelling human saliency gives the GTS model an advantage over the feature-design based unsupervised learning approaches. Comparing with eSDC_knn and eSDC_ocsvm, which are also patch based unsupervised saliency learning methods, the GTS model still shows a notable improvement, e.g. Rank-1 5% (VIPeR) and 15% (iLIDS) relative improvement over eSDC_ocsvm. Figure 4.2 sheds some light into why the GTS model outperforms these two models in [33]. It is evident that a better saliency map is obtained using the GTS model. This is mainly because our topic model explicitly models human appearance as well as background so that the background cannot be mistaken as distractions to true foreground local salient region discovery. In contrast, the model of [33] can give false high saliency scores due to confusion with background regions, while the saliency scores for those real salient regions on those image are pulled down due to the interference of backgrounds, thus cannot be utilised in the re-id process. Computationally, the GTS model is also twice as fast to compute when compared to [33].

4.5.2 Supervised Competitors Evaluation

We also compared our GTS model against some recently proposed supervised learning based re-id models. In general, supervised learning of discriminative models are expected to provide better re-id performance due to the use of labelled information for learning strong discriminative functions, with a high price for labelling the data. Tables 4.1 and 4.2 show results on VIPeR and iLIDS respectively. It is clear that without using any labelled data for model training, the GTS model is competitive against these supervised learning methods *without* the benefit from

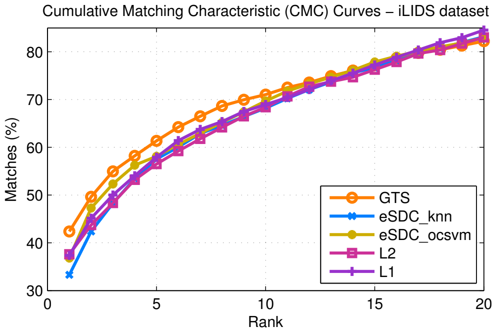


Figure 4.6: iLIDS test: CMC comparison of unsupervised learning based re-id models.

Method	r=1	r=5	r=10	r=20
SDC_knn	33.31	57.55	68.22	83.13
SDC_ocsvm	36.81	58.10	69.69	82.94
PRDC	37.83	63.70	75.09	88.35
LMNN	27.97	53.75	66.14	82.33
PLS	22.10	46.04	59.95	78.68
ITM	28.96	53.99	70.50	86.67
GTS	42.39	61.35	71.04	82.21

Table 4.2: iLIDS test: Comparing the GTS model against other unsupervised (top) and supervised (bottom) learning based models.

learning strong discriminative functions using labelled data. Moreover, the GTS model is able to outperform a number of the supervised learning models by some notable margins, e.g. Rank-1 20% (VIPeR) and 13% (iLIDS) relative improvement over PRDC, LMNN and KISSME (Tables 4.1 and 4.2). This suggests that the GTS model is scalable to large scale applications when manual annotations of identity labels across camera views are not available or feasible.

4.6 Summary

We proposed a novel unsupervised generative saliency learning framework for person re-identification. The core of this framework is a probabilistic topic model specifically designed for modelling jointly typical human appearance and the surrounding background appearance. The model can be deployed to simultaneously learn a saliency map and foreground segmentation for a more accurate and scalable person re-identification model. Compared with existing unsupervised learning methods, the GTS model improves re-id accuracy significantly, especially on Rank-1. The GTS model is also competitive against a couple of supervised learning based competitors, but without requiring manual labelling of data, resulting in greater scalability to large scale re-id problems in many practical applications.

Chapter 5

Incremental Learning from Actively Labelled Data

5.1 Overview

Training unsupervised re-identification models from *only* unlabelled data (Chapter 3, 4) is appealing since it does not require any forms of human annotation. However, a trade-off to these models often sacrifice discriminative power, and are less effective in re-id matching performance compared to the fully-supervised models trained with extensively labelled data. One question arises: can we exploit the advantage of both unsupervised and fully-supervised learning, so that model scalability and discriminative capability can be achieved simultaneously? We notice that real-world re-identification systems often have access to a great extent of unlabelled data, but only afford a very limited human labelling budget for model training purposes. Thus, an interesting problem to investigate regarding to human labelling is that: how to efficiently exploit very few human annotations, but to learn a most discriminative re-id model? Specifically, existing methods are limited due to three reasons:

1. *Small Sample Size*: Due to the high labelling costs and the limited human labour budget, the labelled training population is small in number compared to the searching space at deployment stage. Moreover, the available image samples for each person in typical re-id training data is very limited, e.g. one-shot or a few shots. The SSS problem can thus result in singular intra-class and poor inter-class scatter estimations, indications of problematic class distributions, which in turn lead to suboptimal discriminative solutions.

2. *Unselective Data Labelling*: Given the limited labelling budget and the small training data size, it is thus important to be selective when choosing which data to label among the vast quantity of unlabelled person images generated by a surveillance camera network. In other words, labelling efforts should only be spent on those most representative data samples which contribute most to a model's discriminative capability. However, existing approaches only assume random data selection for labelling, which is a waste of resource.

3. *Offline Labelling*: Existing person re-id methods usually consider off-line batch-wise model learning. However, real-world data collection is an incremental procedure. That is, additional labelled images are available for model training over time, instead of being collected together once at the same time. Also, both camera viewing conditions and population appearance patterns may vary over time. It is thus highly desirable for a re-id model to grow and adapt continuously to the increasingly available labelled data. Moreover, a continuously improving re-id model with an increasingly accurate matching performance will also make the labelling work progressively easier. Given the existing re-id models, this can only be achieved by re-training a model from scratch, resulting in not only high computational cost but also slow response time to a user. They are thus unsuitable for any human-in-the-loop model adaptation.

In this chapter, these three limitations are addressed by formulating person re-id as a *regression* problem [119] with active learning strategies. In particular, the proposed approach has several advantages over existing methods: (1) It has a very simple and efficient closed-form solution with only linear equations; (2) It does not aim to model intra-person variation/distribution thus can accommodate arbitrary sample size per person, e.g. one-shot, effectively mitigating the small sample size problem in re-id; (3) It is readily extended to incremental learning, enabling real-time online model update to incorporate newly available data from model deployment; (4) Its incremental capability can facilitate active sampling to minimise data annotation effort and maximise labelling cost-effectiveness. The **contributions** of this chapter are:

1. We formulate person re-id as an identity regression embedding problem, designed to better cope with the small sample size problem inherent to person re-id. This is in contrast to all existing methods that aim to learn either a classification, verification, or ranking embedding space, which all suffer from the small sample size problem. In particular, we construct explicitly an *identity regression space* defined by the different person identities

of the training population, with each dimension uniquely representing each training person class.

2. We introduce an Identity Regression Space (IRS) model for learning a regression function that maps the raw image feature space to the identity regression space. This IRS model is extremely efficient to compute due to its closed-form solution. This is in contrast with existing classification, verification and ranking based models which need to solve a generalised eigen-problem or some expensive iterative optimisation.

3. We extend the proposed IRS model for incremental learning by deriving an on-line model update algorithm. Instead of learning from scratch for each model update as required by most existing methods, this IRS incremental learning model enables to rapidly build a re-id model from piecewise new data *only*, and progressively adapt the model to more data when available.

4. We further introduce a new active learning algorithm for cost-effective human-in-the-loop incremental model learning and update, by only querying the most informative rather than randomly sampled feedback from a human operator. This active learning model aims to jointly explore the population diversity and discover the class boundary of the up-to-date model. This is mostly lacking in all existing person re-id models.

Extensive experiments on four benchmark datasets VIPeR [7], CUHK01 [108], CUHK03 [1] and Market-1501 [2] demonstrate the superiority and advantages of the proposed IRS model over a wide range of state-of-the-art person re-id models. [21], Additional evaluation and analysis are given to validate the efficacy of the proposed incremental learning and active sampling algorithms for on-line model adaptation.

In the following sections, first we introduce our basic IRS model in the context of a standard supervised learning scenario, whereas its incremental extension and its active learning algorithm are proposed in later sections.

5.2 Identity Regression Learning

5.2.1 Problem Definition

We first consider the image-based person re-identification (re-id) problem [27] in a standard supervised learning setting to introduce our base matching model. The key is to handle the uncontrolled and complex person appearance variations caused by the significant discrepancy in camera viewing condition and human behavioural pose. Similar to existing supervised learning based re-id approaches, we aim to formulate a discriminative feature embedding model capable of effectively and efficiently revealing identity related information of person images from different camera views.

Formally, we assume a labelled training dataset $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ where $\boldsymbol{x}_i \in \mathbb{R}^{d \times 1}$ denotes the d-dimensional feature vector of image $\boldsymbol{x}_i$, with the corresponding identity label vector $\boldsymbol{l} = [l_1, \cdots, l_i, \cdots, l_n] \in \mathbb{Z}^{1 \times n}$, where $l_i \in \{1, \cdots, c\}$ represents the identity label of image $\boldsymbol{x}_i$ among a total of c identities. So, these n training images describe c different persons captured under multiple camera views. We omit the camera label here for brevity. The model learning objective is to obtain a discriminative feature embedding $\boldsymbol{P} \in \mathbb{R}^{d \times m}$, i.e. in the embedding space, the distance between intra-person images is small whilst that of inter-person images is large regardless of their source camera views. In most existing works, the above criterion of compressing intra-person distributions and expanding inter-person distributions is encoded as classification / verification / ranking losses and then a feature embedding is learned by optimising the corresponding objective formulation. However, due to the *Small Sample Size* problem, the learned embedding space is often suboptimal and less discriminative. In addition, there is often no clear interpretation on the learned embedding space.

Our method is significantly different: Prior to the model training, we first explicitly define an *ideal embedding space*, and then train a regression from the raw feature space to the defined embedding space. The learned regression function is our discriminative feature embedding. Specifically, we define a set of "*ideal*" target vectors in the embedding space, denoted by $\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_n^\top]^\top \in \mathbb{R}^{n \times m}$, and explicitly assign them to each of the training sample $\boldsymbol{x}_i$, with $\boldsymbol{y}_i \in \mathbb{R}^{1 \times m}$ referring to $\boldsymbol{x}_i$'s target point in the embedding space, $i \in \{1, 2, \cdots, n\}$ and m referring to the embedding space dimension. In model training, we aim to obtain an optimal feature embedding $\boldsymbol{P}$ that transforms the image feature $\boldsymbol{x}$ into its mapping $\boldsymbol{y}$ with labelled training data $\boldsymbol{X}$. During model deployment, given a test probe image $\tilde{\boldsymbol{x}}^p$ and a set of test gallery images $\{\tilde{\boldsymbol{x}}_i^g\}$, we first transform them into the embedding space with the learned feature embedding $\boldsymbol{P}$, denoted as $\tilde{\boldsymbol{y}}^p$ and $\{\tilde{\boldsymbol{y}}_i^g\}$ respectively. Then, we compute the pairwise matching distances between $\tilde{\boldsymbol{y}}^p$ and $\{\tilde{\boldsymbol{y}}_i^g\}$ by the Euclidean metric. Based on matching distances, we rank all gallery images in ascendant order. Ideally, the true match of the probe person is supposed to appear among top ranks.

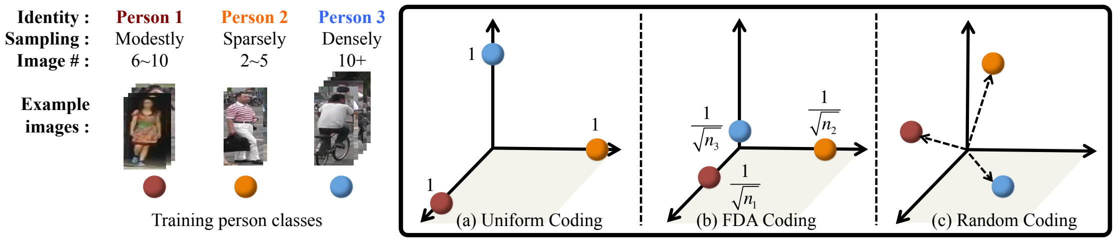


Figure 5.1: Illustration of embedding spaces obtained by three training coding methods. Note, n_i in (b) refers to the training image number of person i extracted from any cameras.

5.2.2 Identity Regression Space

To learn an optimal regression function as feature embedding, one key question in our framework is how to design the target "*ideal*" embedding space, in other words, how to set $\boldsymbol{Y}$. We consider two principles in designing distribution patterns of training samples in the embedding space:

1. *Compactness:* This principle concerns image samples belonging to the *same person class*. Even though each person's intra-class distributions may be different in the raw feature space, we argue that in an optimal embedding space for re-id, the variance of all intra-class distributions should be suppressed. Specifically, for every training person, regardless of the corresponding sample size, all samples should be collapsed to a single point so that the embedding space becomes maximally discriminative with respect to person identity.

2. *Separateness:* This principle concerns image samples belonging to the *different person classes*. Intuitively, the points of different person identities should be maximally separated in the embedding space. With a more intuitive geometry explanation, these points should be located on the vertices of a regular simplex with equal-length edges, so that the embedding space treats equally any training person with a well-separated symmetric structure.

Formally, we assign a unit-length vector on each dimension axis in the embedding space to every training person identity, i.e. we set $\boldsymbol{y}_i = [y_{i,1}, \cdots, y_{i,m}]$ for the i-th training person (Figure 5.1(a)) as:

$$y_{i,j} = \begin{cases} 1, & \text{if} \quad l_i = j; \\ 0, & \text{if} \quad l_i \neq j. \end{cases} \quad \text{with} \quad j \in [1, 2, \cdots, m], \tag{5.1}$$

where l_i is the identity label of image $\boldsymbol{x}_i$. We name this way of setting $\boldsymbol{Y}$ as *Uniform Coding*. The embedding space defined by Eq. (5.1) has a few interesting properties:

1. Each dimension in the embedding space corresponds to one specific training person's identity;

2. Training persons are evenly distributed in the embedding space and the distances between any two training persons are identical;

3. Geometrically, the points of all training person identities together form a standard simplex.

Because each dimension of this embedding space can be now interpreted by one specific training identity, we call such an embedding space an *identity regression space*. Having the identity regression space defined by Eq. (5.1), we propose to exploit the multivariate ridge regression algorithm [119, 120]. In particular, by treating $\boldsymbol{Y}$ as the regression output and $\boldsymbol{P}$ as the to-be-learned parameter, we search for a discriminative projection by minimising the least mean squared error as:

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} \frac{1}{2}\|\boldsymbol{X}^\top \boldsymbol{P} - \boldsymbol{Y}\|_F^2 + \lambda \|\boldsymbol{P}\|_F^2, \tag{5.2}$$

where $\|\cdot\|_F$ is the Frobenius norm, λ controls the regularisation strength. Critically, this formulation has an efficient closed-form solution [119]:

$$\boldsymbol{P}^* = \big(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}\big)^\dagger \boldsymbol{X}\boldsymbol{Y}, \tag{5.3}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose inverse, and $\boldsymbol{I}$ the identity matrix. Since our model learning is by regression towards a training identity space, we call this method the "Identity Regression Space" (**IRS**) model (Figure 5.2).

Discussion. We further discuss the proposed IRS model on the following three aspects:

1. It does not need to calculate any within-class scatters or estimate intra-person distributions, thus it is well suited for mitigating the SSS problem;

2. Compared to most existing methods, the *compactness* criterion can be viewed as an extreme case of minimising the intra-class scatter as in LDA [118] for obtaining better embedding, enjoying a similar spirit and advantage as [47]; and

3. Our *separateness* criterion differs significantly from the conventional way of achieving a discriminative embedding space by learning from the inter-class scatter for separating distinct person classes. Specifically, by treating all training persons equally and distributing

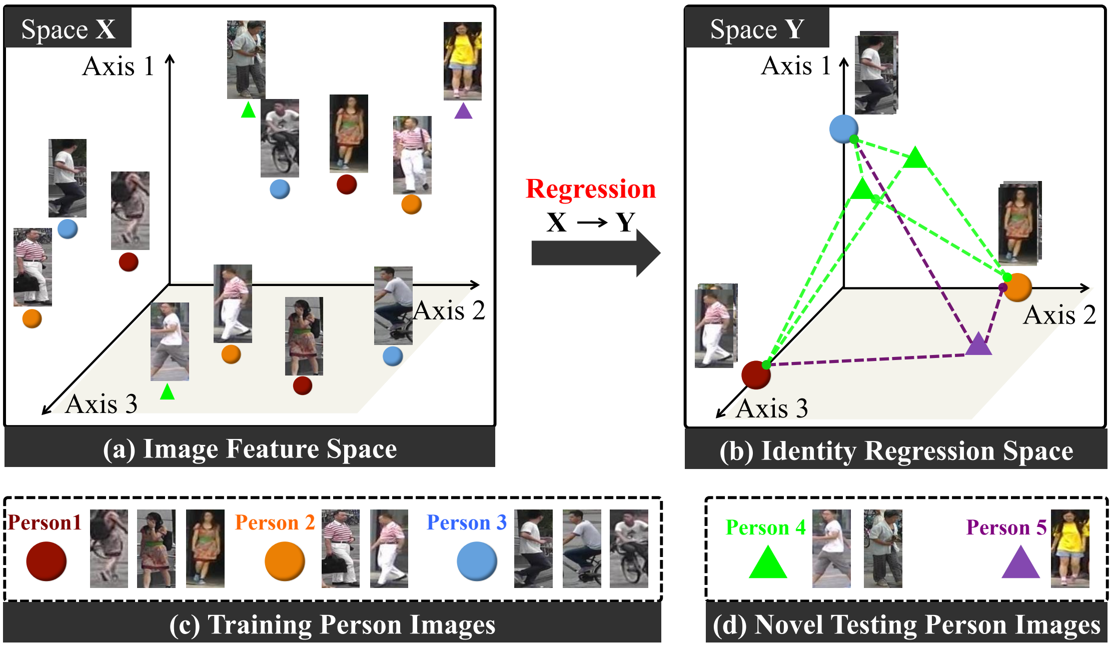


Figure 5.2: Illustration of our Identity Regression Space (IRS) person re-id model. During model training, by regression we learn an identity discriminative feature embedding from (a) the image feature space to (b) the proposed identity regression space defined by (c) all training person classes (indicated by circles). During deployment, we can exploit the learned feature embedding to re-identify (d) novel testing person identities (indicated by triangles) in IRS.

them evenly in the embedding space, the learned feature embedding may be better generalisable to previously unseen testing population as compared to existing methods that take the learning-to-optimise principle without guarantee to induce such a regular embedding space as the IRS model.

Alternative Coding. Apart from the above Uniform Coding (Eq. (5.1)), other designs of the embedding space can also be readily incorporated into our IRS model. We consider two alternative coding methods. The first approach respects the Fisher Discriminant Analysis (FDA) [121, 122] criterion, named *FDA Coding*, which is adopted in the preliminary version of this work [95]. Formally, the FDA criterion can be encoded into our IRS model by setting target identity regression space as (Figure 5.1(b)):

$$y_{ij} = \begin{cases} \frac{1}{\sqrt{n_i}}, & \text{if} \quad l_i = j; \\ 0, & \text{if} \quad l_i \neq j. \end{cases} \quad \text{with} \quad j \in [1, 2, \cdots, m]. \tag{5.4}$$

where n_i and l_i refers to the total image number and identity label of training person i. A detailed derivation is provided in Appendix A. As opposite to Eq. (5.1) which treats each person identity equally (e.g. assigning them with unit-length vectors in the embedding space), this FDA coding scheme assigns variable-length vectors with the length determined by n_i. As shown in (Figure 5.1(b)), with the FDA criterion, the resulting training identity simplex in the embedding space is no longer regular. This may bring benefits for typical classification problems by making size-sensitive use of available training data for modelling individual classes as well as possible, but not necessarily for re-id. Particularly, modelling training classes in such a biased way may instead hurt the overall performance since the re-id model is differently required to generalise the knowledge from seen training person classes to completely unseen testing ones other than within the training ones as in common classification problems.

The second alternative is *Random Coding*. That is, we allocate for each training identity a m-dimensional random vector with every element following a uniform distribution over the range of $[0,1]$ (Figure 5.1(c)), which has shown encouraging effect in shape retrieval [123] and face recognition [124]. In this way, individual dimensions are no longer identity-specific and training identity regression space are shared largely irregularly. We will evaluate the effectiveness of these three coding methods in Section 5.6.1.

Kernalisation. Given complex variations in viewing condition across cameras, the optimal subspace may not be obtainable by linear projections. Therefore, we further kernelise the IRS model (Eq. (5.3)) by projecting the data from the original visual feature space into a reproducing kernel Hilbert space $\mathcal{H}$ with an implicit feature mapping function $\phi(\cdot)$. The inner-product of two data points in $\mathcal{H}$ can be computed by a kernel function: $h_{\text{k}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. By h_{k} (we utilised the typical RBF or Gaussian kernel in our implementation), we obtain a kernel representation $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, based on which a corresponding non-linear projection solution can be induced as:

$$\boldsymbol{Q}^* = \big(\boldsymbol{K}\boldsymbol{K}^\top + \lambda \boldsymbol{K}\big)^\dagger \boldsymbol{K}\boldsymbol{Y}. \tag{5.5}$$

During deployment, different from the linear case, all test samples need to be transformed into the kernel space with h_k before applying the learned projection $\boldsymbol{Q}^*$.

5.3 Incremental Identity Regression Learning

In Section 5.2, we presented the proposed IRS person re-id model. Similar to the majority of conventional re-id methods, we assume a batch-wise model learning setting: First collecting all labelled training data and then learning the feature embedding model (Figure 5.3 (a)). In real-

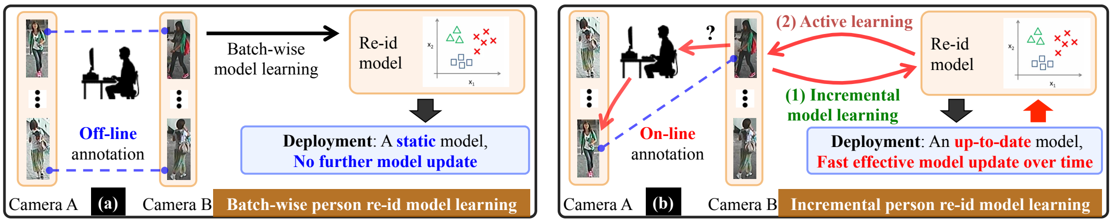


Figure 5.3: Illustration of different person re-id model learning settings. **(a)** Batch-wise person re-id model learning: A re-id model is first learned on an exhaustively labelled training set, and then fixed for deployment without model update; **(b)** Incremental person re-id model learning: Training samples are collected sequentially on-the-fly with either random or active unlabelled data selection, and the re-id model keeps up-to-date by efficient incremental learning from the newly labelled data over time.

world scenario, however, data annotation is likely to arrive in sequence rather than at one time. In such case, a practical system requires the incremental learning capability for cumulatively learning and updating the re-id model over deployment process (Figure 5.3 (b)-(1)). On the other hand, incremental learning is essential for temporal model adaptation, e.g. handling the dynamics in the deployment context [86]. A simple and straightforward scheme is to re-train the model from scratch using the entire training dataset whenever any newly labelled samples become available. Obviously, this is neither computational friendly nor scalable particularly for resource restricted deployment such as on mobile devices.

To overcome this limitation, we introduce an incremental learning algorithm, named IRS$^{\text{inc}}$, for enabling fast model update without the need for re-training from scratch. Suppose at time t, we have the feature matrix $\boldsymbol{X}_t \in \mathbb{R}^{d \times n_t}$ of n_t previously labelled images of c_t person identities, along with $\boldsymbol{Y}_t \in \mathbb{R}^{n_t \times m}$ their indicator matrix defined by Eq. (5.1). We also have the feature matrix $\boldsymbol{X}' \in \mathbb{R}^{d \times n'}$ of n' newly labelled images of c' new person classes, with $\boldsymbol{Y}' \in \mathbb{R}^{n' \times (c_t + c')}$ the corresponding indicator matrix similarly defined by Eq. (5.1). After merging the new data, the updated feature and identity embedding matrix can be represented as:

$$\boldsymbol{X}_{t+1} = [\boldsymbol{X}_t, \boldsymbol{X}'], \quad \boldsymbol{Y}_{t+1} = \begin{bmatrix} \boldsymbol{Y}_t \oplus \boldsymbol{0} \\ \boldsymbol{Y}' \end{bmatrix}, \tag{5.6}$$

where $(\cdot) \oplus \boldsymbol{0}$ denotes the matrix augmentation operation, i.e. padding an appropriate number of zero columns on the right. By defining

$$\boldsymbol{T}_t = \boldsymbol{X}_t \boldsymbol{X}_t^\top + \lambda \boldsymbol{I}, \tag{5.7}$$

and applying Eq. (5.6), we have

$$\boldsymbol{T}_{t+1} = \boldsymbol{T}_t + \boldsymbol{X}'\boldsymbol{X}'^{\top}. \tag{5.8}$$

Also, we can express the projection $\boldsymbol{P}_t \in \mathbb{R}^{d\times m}$ (Eq. (5.3)) of our IRS model at time t as

$$\boldsymbol{P}_t = \boldsymbol{T}_t^{\dagger}\boldsymbol{X}_t\boldsymbol{Y}_t. \tag{5.9}$$

Our aim is to obtain the feature embedding $\boldsymbol{P}_{t+1}$, which requires to compute $\boldsymbol{T}_{t+1}^{\dagger}$. This can be achieved by applying the Sherman-Morrison-Woodbury formula [125] to Eq. (5.8) as:

$$\boldsymbol{T}_{t+1}^{\dagger} = \boldsymbol{T}_t^{\dagger} - \boldsymbol{T}_t^{\dagger}\boldsymbol{X}'\big(\boldsymbol{I} + \boldsymbol{X}'^{\top}\boldsymbol{T}_t^{\dagger}\boldsymbol{X}'\big)^{\dagger}\boldsymbol{X}'^{\top}\boldsymbol{T}_t^{\dagger}. \tag{5.10}$$

Eq. (5.3) and Eq. (5.6) together give us:

$$\begin{aligned}\boldsymbol{P}_{t+1} &= \boldsymbol{T}_{t+1}^{\dagger}\boldsymbol{X}_{t+1}\boldsymbol{Y}_{t+1} \\ &= (\boldsymbol{T}_{t+1}^{\dagger}\boldsymbol{X}_t\boldsymbol{Y}_t) \oplus \boldsymbol{0} + \boldsymbol{T}_{t+1}^{\dagger}\boldsymbol{X}'\boldsymbol{Y}'.\end{aligned} \tag{5.11}$$

Further with Eq. (5.10) and Eq. (5.9), we can update $\boldsymbol{P}$ as:

$$\begin{aligned}\boldsymbol{P}_{t+1} = \Big(\boldsymbol{P}_t - \boldsymbol{T}_t^{\dagger}\boldsymbol{X}'\big(\boldsymbol{I} + \boldsymbol{X}'^{\top}\boldsymbol{T}_t^{\dagger}\boldsymbol{X}'\big)^{\dagger}\boldsymbol{X}'^{\top}\boldsymbol{P}_t\Big) \oplus \boldsymbol{0} \\ + \boldsymbol{T}_{t+1}^{\dagger}\boldsymbol{X}'\boldsymbol{Y}'.\end{aligned} \tag{5.12}$$

Note that, our model update (Eq. (5.10) and Eq. (5.12)) only involves newly coming data samples. As a result, our method does not require to store the training data once utilised for model update. As only cheap computational cost is involved in such linear operations, this proposed algorithm well suits for on-line re-id model learning and updating over the deployment process.

Implementation Consideration. Our IRS$^{\text{inc}}$ model supports incremental learning given either a single new sample ($n' = 1$) or a small chunk of samples ($n' \geqslant 2$). If the data chunk size $n' \ll d$ (where d is the feature dimension), it is faster to perform n' separate updates on each new sample instead of by a whole chunk. The reason is that, in such a way the Moore-Penrose matrix inverse in Eq. (5.10) and Eq. (5.12) can be reduced to n' separate scaler inverses, which is much cheaper in numerical computation.

5.4 Active Identity Regression Learning

The incremental learning process described above is *passive*, i.e. a human annotator is supposed to label randomly chosen data without considering the potential value of each selected sample in improving the re-id model. Therefore, data annotation by this random way is likely to contain redundant information with partial labelling effort wasted. To resolve this problem, we explore the active learning idea [71] for obtaining more cost-effective incremental re-id model update (Figure 5.3 (b)-(2)).

Active IRS$^{\text{inc}}$ Overview. In practice, we often have access to a large number of *unlabelled* images $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{G}}$ captured by disjoint cameras. Assume at time step $t \in \{1, \cdots, \tau\}$ with τ defining the pre-determined human labelling budget, we have the up-to-date IRS$^{\text{inc}}$ model m_t (corresponding to the feature embedding $\boldsymbol{P}_t$), along with $\widetilde{\mathcal{P}}_t$ and $\widetilde{\mathcal{G}}_t$ denoting the remaining unlabelled data. To maximise labelling profit, we propose an *active labelling* algorithm for IRS$^{\text{inc}}$ with the main steps as follows:

1. An image $\boldsymbol{x}_t^p \in \widetilde{\mathcal{P}}_t$ of a new training identity l_t is *actively* selected by model m_t, according to its potential usefulness and importance measured by certain active sampling criteria (see details below);

2. A ranking list of unlabelled images $\widetilde{\mathcal{G}}_t$ against the selected $\boldsymbol{x}_t^p$ is then generated by m_t based matching distances;

3. For the selected $\boldsymbol{x}_t^p$, a human annotator is then asked to manually identify the cross-view true matching image $\boldsymbol{x}_t^g \in \widetilde{\mathcal{G}}_t$ in the ranking list, and then generate a new annotation ($\boldsymbol{x}_t^p$, $\boldsymbol{x}_t^g$);

4. The IRS$^{\text{inc}}$ re-id model is updated to m_{t+1} (i.e. $\boldsymbol{P}_{t+1}$) from the new data annotation $(\boldsymbol{x}_t^p, \boldsymbol{x}_t^g)$ by our incremental learning algorithm (Eq. (5.10) and Eq. (5.12)).

Among these steps above, the key lies in how to select a good image $\boldsymbol{x}_t^p$. To this end, we derive a “Joint Exploration-Exploitation” (**JointE^2**) active sampling algorithm composed of three criteria as follows (Figure 5.4).

(I) Appearance Diversity Exploration. Intuitively, the appearance diversity of training people is a critical factor for the generalisation capability of a re-id model. Thus, the preferred next image to annotate should lie in the most unexplored region of the population $\widetilde{\mathcal{P}}_t$. Specifically, at

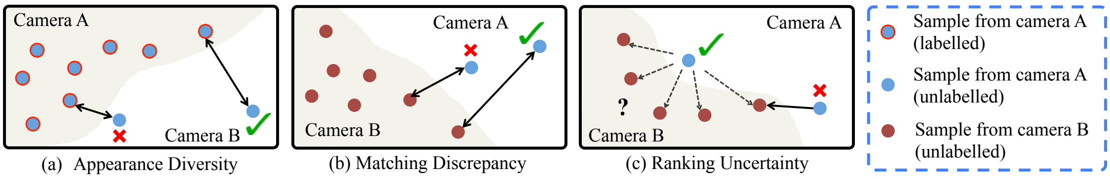


Figure 5.4: Illustration of the proposed active exploration and exploitation selection criteria for more cost-effective incremental re-id model learning.

time t, the distance between any two samples $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ by the current re-id model is computed as:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2 | m_t) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{P}_t \boldsymbol{P}_t^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2). \tag{5.13}$$

Given the unlabelled $\widetilde{\mathcal{P}}_t$ and labelled $\mathcal{P}_t$ part of the set $\widetilde{\mathcal{P}}$ ($\widetilde{\mathcal{P}}_t \bigcup \mathcal{P}_t = \widetilde{\mathcal{P}}$), we can measure the diversity degree of an unlabelled sample $\boldsymbol{x}_i^p \in \widetilde{\mathcal{P}}_t$ by its distance against the *within-view nearest neighbour* in $\mathcal{P}_t$ (Figure 5.4 (a)):

$$\begin{aligned} \varepsilon_1(\boldsymbol{x}_i^p) &= \min\ d(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p | m_t), \\ \text{s.t. } & \boldsymbol{x}_i^p \in \widetilde{\mathcal{P}}_t,\ \boldsymbol{x}_j^p \in \mathcal{P}_t. \end{aligned} \tag{5.14}$$

By doing so, more diverse person appearance can be covered and learned for more rapidly increasing the knowledge of the IRS$^{\text{inc}}$ model, rather than repeatedly learning visually similar training samples.

(II) Matching Discrepancy Exploration. A well learned re-id model is supposed to find the true match of a given image with a small cross-view matching distance. In this perspective, our second criterion particularly prefers the samples with large matching distances in the embedding space, i.e. the re-id model m_t remains largely unclear on what are the likely corresponding cross-view appearances of these "unfamiliar" people. Numerically, we compute the matching distance between an unlabelled sample $\boldsymbol{x}_i^p \in \widetilde{\mathcal{P}}_t$ and the cross-view true match (assumed as *cross-view nearest neighbour*) in $\widetilde{\mathcal{G}}$ (Figure 5.4 (b)):

$$\begin{aligned} \varepsilon_2(\boldsymbol{x}_i^p) &= \min\ d(\boldsymbol{x}_i^p, \boldsymbol{x}_j^g | m_t), \\ \text{s.t. } & \boldsymbol{x}_i^p \in \widetilde{\mathcal{P}}_t,\ \boldsymbol{x}_j^g \in \widetilde{\mathcal{G}}. \end{aligned} \tag{5.15}$$

That is, the unlabelled images with greater $\varepsilon_2(\boldsymbol{x}_i^p)$ are preferred to be selected.

(III) Ranking Uncertainty Exploitation. Uncertainty-based exploitative sampling schemes have been widely investigated for classification problems [126, 127, 81]. The essential idea is to query the least certain sample for human to annotate. Tailored for re-id tasks with this idea, given the similar appearance among different identities, a weak re-id model may probably generate similar ranking scores for those visually ambiguous gallery identities with respect to a given probe. Naturally, it should be useful and informative to manually label such "challenging" samples for enhancing a person re-id model's discrimination power particularly with regarding to such person appearance (Figure 5.4 (c)). To obtain such person images, we define a matching distance based probability distribution over all samples $\boldsymbol{x}_j^g \in \widetilde{\mathcal{G}}$ for a given cross-view image $\boldsymbol{x}_i^p \in \widetilde{\mathcal{P}}$:

$$p_{m_t}(\boldsymbol{x}_j^g | \boldsymbol{x}_i^p) = \frac{1}{Z_i^t} e^{-d(\boldsymbol{x}_i^p, \boldsymbol{x}_j^g | m_t)}, \tag{5.16}$$

where

$$Z_i^t = \sum_k e^{-d(\boldsymbol{x}_i^p, \boldsymbol{x}_k^g | m_t)}, \;\; \boldsymbol{x}_k^g \in \widetilde{\mathcal{G}}.$$

The quantity $p_{m_t}(\boldsymbol{x}_j^g | \boldsymbol{x}_i^p)$ gives a high entropy when most ranking scores are adjacent to each other, indicating great information to mine from the perspective of information theory [128]. In other words, the model has only a low confidence on its generated ranking list considering that only a very few number of cross-camera samples are likely to be true matches rather than many of them. Consequently, our third criterion is designed as:

$$\begin{gathered}\varepsilon_3(\boldsymbol{x}_i^p) = -\sum_j p_{m_t}(\boldsymbol{x}_j^g | \boldsymbol{x}_i^p) \log p_{m_t}(\boldsymbol{x}_j^g | \boldsymbol{x}_i^p), \\ \text{s.t. } \boldsymbol{x}_i^p \in \widetilde{\mathcal{P}}_t, \;\; \boldsymbol{x}_j^g \in \widetilde{\mathcal{G}}.\end{gathered} \tag{5.17}$$

which aims to select out those associated with high model ranking ambiguity.

Joint Exploration-Exploitation. Similar to the model in [79, 81], we combine both exploitation and exploration based criteria into our final active selection standard, formally as:

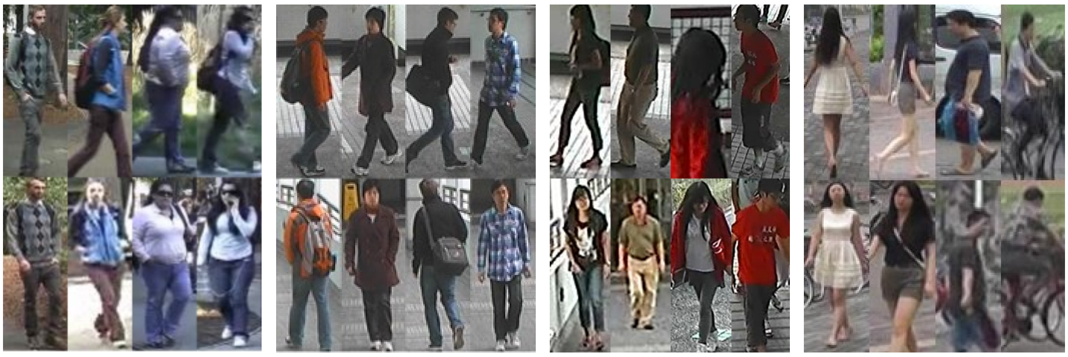
$$\varepsilon(\boldsymbol{x}_i^p) = \varepsilon_1(\boldsymbol{x}_i^p) + \varepsilon_2(\boldsymbol{x}_i^p) + \varepsilon_3(\boldsymbol{x}_i^p). \tag{5.18}$$

Note that, we normalise $\varepsilon_1, \varepsilon_2, \varepsilon_3$ to the unit range $[0, 1]$ respectively before performing this fusion for eliminating the scale discrepancy problem.

In summary, with Eq. (5.18), all the unlabelled samples in $\widetilde{\mathcal{P}}$ can be sorted according, and the one with highest $\varepsilon(\boldsymbol{x}_i^p)$ is then selected for human annotation. An overview of our proposed

Algorithm 1: Active IRS$^{\text{inc}}$

```
Data:
    (1) Unlabelled image set $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{G}}$ from disjoint cameras;
    (2) Regularisation strength $\lambda$;
    (3) Labelling budget $\tau$.
Result:
    (1) Discriminative feature embedding matrix $\boldsymbol{P}$;
Initialisation:
    (1) Randomly label a small seed set $\boldsymbol{X}_0$, $\boldsymbol{Y}_0$;
    (2) Set $\boldsymbol{T}_0^{\dagger} = (\boldsymbol{X}_0\boldsymbol{X}_0^{\top} + \lambda\boldsymbol{I})^{\dagger}$;
    (3) Set $\boldsymbol{P}_0 = \boldsymbol{T}_0^{\dagger}\boldsymbol{X}_0\boldsymbol{Y}_0$ (Eq. (5.3)).
Active Labelling:
for $t = 0 : \tau - 1$ do
    (1) Select an unlabelled sample $\boldsymbol{x}_t^p \in \widetilde{\mathcal{P}}_t$ (Eq. (5.18));
    (2) Rank the images in $\widetilde{\mathcal{G}}_t$ against the selection $\boldsymbol{x}_t^p$;
    (3) Human annotator verifies the true match in $\widetilde{\mathcal{G}}_t$;
    (4) Generate a new annotation $(\mathcal{I}_t^p, \mathcal{I}_t^g)$;
    (5) Update $\boldsymbol{T}_{t+1}^{\dagger}$ (Eq. (5.10));
    (6) Update $\boldsymbol{P}_{t+1}$ (Eq. (5.12)).
return $\boldsymbol{P} = \boldsymbol{P}_{\tau}$;
```



(a) VIPeR (b) CUHK01 (c) CUHK03 (d) Market-1501

Figure 5.5: Example person images from four person re-id datasets. Two images of the same column describe the same person.

active learning based incremental model learning and updating is presented in Algorithm 13. We will show the effect of our proposed active labelling method in our evaluations (Section 5.6.2).

5.5 Datasets and Experimental Settings

Datasets. For model evaluation, four person re-id benchmarks were used: VIPeR [7], CUHK01 [108], CUHK03 [1], and Market-1501 [2], as summarised in Table 5.1. We show in Figure 5.5 some examples of person images from these datasets. Note that the datasets were collected with different data sampling protocols: (a) VIPeR has one image per person per view; (b) CUHK01

Dataset	Cameras	Persons	Labelled BBox	Detected BBox
VIPeR	2	632	1,264	0
CUHK01	2	971	1,942	0
CUHK03	6	1,467	14,097	14,097
Market-1501	6	1,501	0	32,668

Table 5.1: Statistics of person re-id datasets. BBox: Bounding Box.

contains two images person per view; (c) CUHK03 consists of a maximum of five images per person per view, and also provides both manually labelled and auto-detected image bounding boxes with the latter posing more challenging re-id test due to unknown misalignment of the detected bounding boxes; (d) Market-1501 has variable numbers of images per person per view. These four datasets present a good selection of re-id test scenarios with different population sizes under realistic viewing conditions exposed to large variations in human pose and strong similarities among different people.

Features. To capture the detailed information of person appearance, we adopted three state-of-the-art feature representations with variable dimensionalities from 10^4 to 10^2: **(1)** *Local Maximal Occurrence* (LOMO) feature [24]: The LOMO feature is based on a HSV colour histogram and Scale Invariant Local Ternary Pattern [129]. For alleviating the negative effects caused by camera view discrepancy, the Retinex algorithm [130] is applied to pre-process person images. The feature dimension of LOMO is rather high at $26,960$, therefore expensive to compute. **(2)** *Weighted Histograms of Overlapping Stripes* (WHOS) feature [61, 131]: The WHOS feature contains HS/RGB histograms and HOG [29] of image grids, with a centre support kernel as weighting to approximately segmented person foreground from background clutters. We implemented this feature model as described by [61]. The feature dimension of WHOS is moderate at $5,138$. **(3)** *Convolutional Neural Network* (CNN) feature [9]: Unlike hand-crafted LOMO and WHOS features, deep CNN person features are learned from image data. Specifically, we adopted the CNN model of [9] and used the FC_7 layer output as the deep feature for person re-id. This CNN FC_7 feature has a rather low dimension of 256, thus easy to compute. To compute this deep feature, we first trained the CNN model with authors' released code on the $26,246$ training images of CUHK03.We then deployed the trained CNN model to extract features of the test data of CUHK03 (same domain). On Market-1501, the CUHK03 trained CNN was further fine-tuned using the $12,936$ training person images of Market-1501 for feature domain adaptation. On VIPeR and CUHK01, the CUHK03 trained CNN was *directly* deployed *without* any fine-tuning as there are insufficient training images to make effective deep feature domain adaptation, with only 632 and $1,940$ training images for VIPeR and CUHK01 respectively.

Model Training Settings. In evaluations, we considered extensively comparative experiments under two person re-id model training settings: **(I)** *Batch-wise model training*: In this setting, we followed the conventional supervised re-id scheme commonly utilised in most existing methods, that is, first collecting all training data and then learning a re-id model *before* deployment. **(II)** *Incremental model training*: In contrast to the batch-wise learning, we further evaluated a more realistic data labelling scenario where more training labels are further collected over time *after* model deployment. The proposed IRS$^{\text{inc}}$ model was deployed for this incremental learning setting.

5.6 Experiments and Evaluations

5.6.1 Batch-Wise Person Re-Identification Evaluation

Batch-Wise Re-Id Evaluation Protocol. To facilitate quantitative comparisons with existing re-id methods, we adopted the standard supervised re-id setting to evaluate the proposed IRS model. Specifically, on *VIPeR*, we split randomly the whole population of the dataset (632 people) into two halves: One for training (316) and another for testing (316). We repeated 10 trials of random people splits and utilised the averaged results. On *CUHK01*, we considered two benchmarking training/test people split settings: (1) 485/486 split: randomly selecting 485 identities for training and the other 486 for testing [24, 47]; (2) 871/100 split: randomly selecting 871 identities for training and the other 100 for testing [50, 51]. As CUHK01 is a multi-shot (e.g. multiple images per person per camera view) dataset, we computed the final matching distance between two people by averaging corresponding cross-view image pairs. Again, we reported the results averaged over 10 random trials for either people split. On *CUHK03*, following [1] we repeated 20 times of random 1260/100 people splits for model training/test and reported the averaged accuracies under the single-shot evaluation setting[47]. On *Market-1501*, we used the standard training/test (750/751) people split provided by [2]. On all datasets, we exploited the cumulative matching characteristic (CMC) to measure the re-id accuracy performance. On Market-1501, we also considered the recall measure of multiple truth matches by mean Average Precision (mAP), i.e. first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes [2].

In the followings, we evaluated: (i) Comparative person re-id performance of our IRS model against existing state-of-the-art methods, (ii) the effects of different embedding spaces on the IRS model, (iii) the effects of feature choices on the IRS model, (iv) the sensitivity of parameter λ in Eq. (5.2), and (v) model complexity and computational costs among different methods.

(I) Comparisons to the State-of-The-Art. We first evaluated the proposed IRS model by extensive comparisons to the existing state-of-the-art re-id models under the standard supervised person re-id setting. We considered a wide range of existing re-id methods, including both hand-crafted and deep learning models. In the following experiments, we deployed the *Uniform Coding* (Eq. (5.1) in Section 5.2.2) for the identity regression space embedding of our IRS model unless stated otherwise. We considered both single- and multi-feature based person re-id performance, and also compared re-id performances of different models on auto-detected bounding boxes when available in CUHK03 and Market-1501.

Evaluation on VIPeR. Table 5.2 shows a comprehensive comparison on re-id performance between our IRS model (and its variations) and 43 existing models using the VIPeR benchmark [7]. It is evident that our IRS model with a non-deep feature LOMO, IRS(LOMO), is better than all existing methods[1] except the deep model MCP [57], with Rank-1 45.1% vs. 47.5% respectively. Interestingly, using our CUHK03 trained CNN deep feature *without* fine-tuning on VIPeR, i.e. IRS(CNN), does not offer extra advantage (Rank-1 33.1%), due to the significant domain drift between VIPeR and CUHK03. This becomes more clear when compared with the CUHK01 tests below. Moreover, given a score-level fusion on the matching of three different features, IRS(WHOS+LOMO+CNN), the IRS model can benefit from further boosting on its re-id performance, obtaining the best Rank-1 rate at 54.6%. These results demonstrate the effectiveness of the proposed IRS model in learning identity discriminative feature embedding because of our *unique* approach on identity regression to learning a re-id embedding space, in contrast to the existing established ideas on classification, verification or ranking based learning of a re-id model.

Evaluation on CUHK01. Table 5.3 shows a comprehensive comparison of the IRS model with 24 existing re-id models on the CUHK01 benchmark [108]. It is clear that the proposed IRS model achieves the best re-id accuracy under both training/test split protocols. Note that, HER [95] is IRS-FDA(LOMO). Specifically, for the 486/485 split, our IRS(CNN) method surpassed

[1] The HER model presented in our preliminary work [95] is the same as IRS(LOMO) with FDA coding (Eq. (5.4)), i.e. HER = IRS-FDA(LOMO). On the other hand, IRS(LOMO) in Tables 5.2, 5.3, 5.4 and 5.5 is IRS-Uniform(LOMO). The effects of choosing different coding is evaluated later (Table 5.6).

Dataset	VIPeR			
Rank (%)	R1	R5	R10	R20
ISFI [37]	17.1	39.0	52.9	67.3
KISSME [18]	22.0	-	68.0	-
LFDA [20]	24.2	52.0	67.1	82.0
RPLM [34]	27.0	-	69.0	83.0
SalMatch [115]	30.2	52.3	65.5	79.2
MLF [21]	29.1	52.3	66.0	79.9
kLFDA [23]	38.6	69.2	80.4	89.2
SCNCD [45]	33.7	62.7	74.8	85.0
KCCA [61]	37.0	-	85.0	93.0
XQDA [24]	40.0	68.1	80.5	91.1
MLAPG [25]	40.7	69.9	82.3	92.4
RKSL [132]	40.2	74.5	85.7	93.5
NFST [47]	42.3	71.5	82.9	92.1
LSSCDL [49]	42.7	-	84.3	91.9
TMA [86]	43.8	-	83.8	91.5
HER [95]	45.1	74.6	85.1	93.3
DML [62]	28.2	59.3	73.5	86.4
DCNN+ [50]	34.8	63.6	75.6	84.5
RDC-Net[56]	40.5	60.8	70.4	84.4
JRL [133]	38.4	69.2	81.3	90.4
SICI [58]	35.8	-	-	-
DGD [9]	38.6	-	-	
Gated S-CNN [52]	37.8	66.9	77.4	-
EDM [51]	40.9	-	-	-
S-LSTM [63]	42.4	68.7	79.4	-
MCP [57]	**47.8**	74.7	84.8	91.1
IRS (WHOS)	44.5	**75.0**	**86.3**	**93.6**
IRS (LOMO)	45.1	74.6	85.1	93.3
IRS (CNN)	33.1	59.9	71.5	82.2
MLF* [21]	43.4	73.0	84.9	93.7
ME* [46]	45.9	77.5	88.9	95.8
CVDCA* [134]	47.8	76.3	86.3	94.0
FFN-Net* [135]	51.1	81.0	91.4	**96.9**
NFST* [47]	51.2	82.1	90.5	95.9
HER* [95]	53.0	79.8	89.6	95.5
GOG* [96]	49.7	-	88.7	94.5
SCSP* [136]	53.5	**82.6**	**91.5**	96.7
IRS (WHOS+LOMO+CNN)*	**54.6**	81.5	90.3	95.7

Table 5.2: Re-Id performance comparison on the VIPeR benchmark. (*): Multiple features fusion.

the deep learning DGD model [9], the second best in this comparison, by Rank-1 2.0%$(68.6-66.6)$. For the 871/100 split, IRS(CNN) yields a greater performance boost over DGD with improvement on Rank-1 at 12.6%$(84.4-71.8)$. It is also worth pointing out that the DGD model was trained using data from other 6 more datasets and further carefully fine-tuned on CUHK01. In contrast, our IRS(CNN) model was only trained on CUHK03 without fine-tuning on CUHK01, and the CNN architecture we adopted closely resembles to that of DGD. Moreover, by fusing multiple features, the performance margin of IRS(WHOS+LOMO+CNN) over the existing models is further enlarged under both splits, achieving Rank-1 11.7%$(80.8-69.1)$ boost over NFST [47] and Rank-1 16.6%$(88.4-71.8)$ boost over SICI [58], respectively. Compared to VIPeR, the overall re-id performance advantage of the IRS model on CUHK01 is greater over existing models. This is due to not only identity prototype regression based feature embedding,

Dataset	CUHK01 (486/485 split)			
Rank (%)	R1	R5	R10	R20
LMNN [137]	13.4	31.3	42.3	54.1
ITML [138]	16.0	35.2	45.6	59.8
SalMatch [115]	28.5	45.9	55.7	68.0
MLF [21]	20.5	37.1	45.3	55.3
RefDes [139]	31.1	-	68.6	79.2
kLFDA [23]	54.6	80.5	86.9	92.0
CVDCA [134]	47.8	74.2	83.4	89.9
XQDA [24]	63.2	83.9	90.0	94.2
MLAPG [25]	64.2	85.4	90.8	94.9
L_1-Lap [140]	50.1	-	-	-
NFST [47]	65.0	85.0	89.9	94.4
HER [95]	68.3	86.7	92.6	96.2
DCNN+ [50]	47.5	71.6	80.3	87.5
MCP [57]	53.7	84.3	91.0	93.3
DGD [9]	66.6	-	-	-
IRS (WHOS)	48.8	73.4	81.1	88.3
IRS (LOMO)	68.3	86.7	92.6	96.2
IRS (CNN)	**68.6**	**89.3**	**93.9**	**97.2**
ME* [46]	53.4	76.4	84.4	90.5
FFN-Net* [135]	55.5	78.4	83.7	92.6
GOG* [96]	67.3	86.9	91.8	95.9
NFST* [47]	69.1	86.9	91.8	95.4
HER* [95]	71.2	90.0	94.4	97.3
IRS (WHOS+LOMO+CNN)*	**80.8**	**94.6**	**96.9**	**98.7**
Dataset	CUHK01 (871/100 split)			
FPNN [1]	27.9	59.6	73.5	87.3
DCNN+ [50]	65.0	-	-	-
JRL [133]	70.9	92.3	96.9	98.7
EDM [51]	69.4	-	-	-
SICI [58]	71.8	-	-	-
IRS (WHOS)	77.0	92.8	96.5	99.2
IRS (LOMO)	80.3	94.2	96.9	99.5
IRS (CNN)	84.4	98.2	**99.8**	**100**
IRS (WHOS+LOMO+CNN)*	**88.4**	**98.8**	99.6	**100**

Table 5.3: Re-id performance comparison on the CUHK01 benchmark. (*): Multiple features fusion.

but also less domain drift from CUHK03 to CUHK01, given that the CNN feature used by IRS was trained on CUHK03.

Evaluation on CUHK03. The person re-id performance of 19 different methods as compared to the IRS model on CUHK03 [1] is reported in Table 5.4. We tested on both the manually labelled and automatically detected bounding boxes. Similar to VIPeR and CUHK01, our IRS model surpassed clearly all compared methods in either single- or multi-feature setting given manually labelled bounding boxes. Importantly, this advantage remains when more challenging detected bounding boxes were used, whilst other strong models such as NFST and GOG suffered more significant performance degradation. This shows both the robustness of our IRS model against misalignment and its greater scalability to real-world deployments.

Evaluation on Market-1501. We evaluated the re-id performance of 13 existing models against the proposed IRS model on the Market-1501 benchmark [2]. The bounding boxes of all person

Dataset	CUHK03 (Manually)			
Rank (%)	R1	R5	R10	R20
kLFDA [23]	45.8	77.1	86.8	93.1
XQDA [24]	52.2	82.2	92.1	96.3
MLAPG [25]	58.0	87.1	94.7	98.0
NFST [47]	58.9	85.6	92.5	96.3
LSSCDL [49]	57.0	-	-	-
HER [95]	60.8	87.0	95.2	97.7
FPNN [1]	20.7	-	-	-
DCNN+ [50]	54.7	86.5	93.9	98.1
EDM [51]	61.3	-	-	-
DGD [9]	75.3	-	-	
IRS (WHOS)	59.6	87.2	92.8	96.9
IRS (LOMO)	61.6	87.0	94.6	**98.0**
IRS (CNN)	**81.5**	**95.7**	**97.1**	**98.0**
ME* [46]	62.1	89.1	94.3	97.8
NFST* [47]	62.6	90.1	94.8	98.1
HER* [95]	65.2	92.2	96.8	99.1
GOG* [96]	67.3	91.0	96.0	-
IRS (WHOS+LOMO+CNN)*	**81.9**	**96.5**	**98.2**	**98.9**
Dataset	CUHK03 (Detected)			
ITML [138]	5.1	17.7	2.8.3	-
LMNN [137]	6.3	18.7	29.0	-
KISSME [18]	11.7	33.3	48.0	-
BoW [2]	23.0	42.4	52.4	64.2
XQDA [24]	46.3	78.9	83.5	93.2
MLAPG [25]	51.2	83.6	92.1	96.9
L$_1$-Lap [140]	30.4	-	-	-
NFST [47]	53.7	83.1	93.0	94.8
LSSCDL [49]	51.2	80.8	89.6	-
FPNN [1]	19.9	-	-	-
DCNN+ [50]	44.9	76.0	83.5	93.2
EDM [51]	52.0	-	-	-
SICI [58]	52.1	84.9	92.4	-
S-LSTM [63]	57.3	80.1	88.3	-
Gated S-CNN [52]	68.1	88.1	94.6	-
IRS (WHOS)	50.6	82.1	90.4	96.1
IRS (LOMO)	53.4	83.1	91.2	96.4
IRS (CNN)	**80.3**	**96.3**	**98.6**	**99.0**
NFST* [47]	54.7	84.8	94.8	95.2
GOG* [96]	65.5	88.4	93.7	-
IRS (WHOS+LOMO+CNN)*	**83.3**	**96.2**	**97.9**	**98.6**

Table 5.4: Re-id performance comparison on the CUHK03 benchmark. (*): Multiple features fusion.

images of this dataset were generated by an automatic pedestrian detector. Hence, this dataset presents a more realistic challenge to re-id models than conventional re-id datasets with manually labelled bounding boxes. Table 5.5 shows the clear superiority of our IRS model over all competitors. In particular, our IRS model achieved Rank-1 73.9% for single-query and Rank-1 81.4% for multi-query, significantly better than the strongest alternative method, the deep Gated S-CNN model [52], by 8.1%$(73.9-65.8)$ (single-query) and 5.4%$(81.4-76.0)$ (multi-query). Similar advantages hold when compared using the mAP metric.

In summary, these comparative evaluations on the performance of batch-wise re-id model learning show that the IRS model outperforms comprehensively a wide range of existing re-id methods including both hand-crafted and deep learning based models. This validates the effec-

Dataset	Market-1501			
Query Per Person	Single-Query		Multi-Query	
Metric (%)	R1	mAP	R1	mAP
BoW [2]	34.4	14.1	42.6	19.5
KISSME [18]	40.5	19.0	-	-
MFA [141]	45.7	18.2	-	-
kLFDA [23]	51.4	24.4	52.7	27.4
XQDA [24]	43.8	22.2	54.1	28.4
SCSP [136]	51.9	26.3	-	-
NFST [47]	55.4	29.9	68.0	41.9
TMA [86]	47.9	22.3	-	-
HL [142]	59.5	-	-	-
SSDAL [143]	39.4	19.6	49.0	25.8
S-LSTM [63]	-	-	61.6	35.3
Gated S-CNN [52]	65.8	39.5	76.0	48.4
IRS (WHOS)	55.2	27.5	60.3	33.5
IRS (LOMO)	57.7	29.0	68.0	37.8
IRS (CNN)	**72.7**	**48.1**	**80.2**	**58.5**
BoW* [2]	-	-	47.3	21.9
SCSP* [136]	51.9	26.4	-	-
NFST* [47]	61.0	35.7	71.6	46.0
IRS (WHOS+LOMO+CNN)*	**73.9**	**49.4**	**81.4**	**59.9**

Table 5.5: Re-id performance comparison on the Market-1501 benchmark. (*): Multiple features fusion.

Dataset	VIPeR				CUHK01				CUHK03				Market-1501			
Rank (%)	R1	R5	R10	R20	R1	R5	R10	R20	R1	R5	R10	R20	R1(SQ)	mAP(SQ)	R1(MQ)	mAP(MQ)
Uniform Coding	**45.1**	**74.6**	**85.1**	**93.3**	**68.3**	**86.7**	**92.6**	**96.2**	**61.6**	**87.0**	94.6	**98.0**	**57.7**	**29.0**	**68.0**	**37.8**
FDA Coding [95]	**45.1**	**74.6**	**85.1**	**93.3**	**68.3**	**86.7**	**92.6**	**96.2**	60.8	**87.0**	**95.2**	97.7	55.6	27.5	67.5	36.8
Random Coding [123]	44.8	73.4	84.8	92.7	61.3	83.4	89.5	94.2	51.7	79.4	87.4	93.0	47.4	21.1	48.5	23.2

Table 5.6: Effects of embedding space design on person re-id performance in our proposed IRS model. The LOMO visual feature were used on all datasets. We adopted the 485/486 people split on CUHK01 and the manually labelled person images on CUHK03. SQ: Single-Query; MQ: Multi-Query.

tiveness and advantages of learning a re-id discriminative feature embedding using the proposed approach on identity regression.

(II) Effects of Embedding Space Design. To give more insight on why and how the IRS model works, we evaluated the effects of embedding space design in our IRS model. To this end, we compared the three coding methods as described in Section 5.2.2: *Uniform Coding* in the proposed *Identity Regression Space*, *FDA Coding* by [95], and *Random Coding* by [123]. In this experiment, we used the LOMO feature on all four datasets, the 485/486 people split on CUHK01, and the manually labelled bounding boxes on CUHK03. For Random Coding, we performed 10 times and used the averaged results to compare with the Uniform Coding and the FDA Coding. The results are presented in Table 5.6. We have the following observations:

(i) The embedding space choice plays a clear role in IRS re-id model learning and a more "semantic" aligned (both Uniform and FDA) coding has the advantage for learning a more discriminative IRS re-id model. One plausible reason is that the Random Coding may increase the model learning difficulty resulting in an inferior feature embedding, especially given the small sample size nature of re-id model learning. Instead, by explicitly assigning identity class "semantics" (prototypes) to individual dimensions of the embedding space, the feature embedding learning is made more selective and easier to optimise.

(ii) Both the Uniform and FDA Coding methods yield the same re-id accuracy on both VIPeR and CUHK01. This is because on either dataset each training identity has the same number of images (2 for VIPeR and 4 for CUHK01), under which the FDA Coding (Eq. (5.4)) is equivalent to the Uniform Coding (Eq. (5.1)).

(iii) Given the different image samples available per training person identity on CUHK03 and Market-1501, FDA Coding is slightly inferior to Uniform Coding. This is interesting given the robust performance of FDA on conventional classification problems. Our explanation is rather straightforward if one considers the unique characteristics of the re-id problem where the training and test classes are *completely* non-overlapping. That is, the test classes have no training image samples. In essence, the re-id problem is conceptually similar to the problem of Zero-Shot Learning (ZSL), in contrast to the conventional classification problems where test classes are sufficiently represented by the training data, i.e. totally overlapping. More specifically, learning by the FDA criterion optimises a model to the training identity classes given sufficient samples per class but it does not work well with small sample sizes, and more critically, it does *not necessarily* optimise the model for previously unseen test identity classes. This is because if the training identity population is relatively small, as in most re-id datasets, an unseen test person may not be similar to any of training people, That is, the distributions of the training and test population may differ significantly. Without any prior knowledge, a good representation of an unseen test class is some unique combination of all training persons *uniformly* without preference. Therefore, a feature embedding optimised uniformly without bias/weighting by the training class data sampling distribution is more likely to better cope with more diverse and unseen test classes, by better preserving class diversity in the training data *especially given the small sample size challenge* in re-id training data. This can be seen from the regularised properties of the Uniform Coding in Section 5.2.

(III) Effects of Features. We evaluated the effects of three different visual features (WHOS, LOMO, and CNN) individually and also their combinations used in our IRS model with the Uniform Coding, as shown in Table 5.7.

Dataset	VIPeR				CUHK01 (486/485 split)				CUHK01 (871/100 split)			
Metric (%)	R1	R5	R10	R20	R1	R5	R10	R20	R1	R5	R10	R20
WHOS [131]	44.5	**75.0**	**86.3**	**93.6**	48.8	73.4	81.1	88.3	77.0	92.8	96.5	99.2
LOMO [24]	**45.1**	74.6	85.1	93.3	68.3	86.7	92.6	96.2	80.3	94.2	96.9	99.5
CNN [9]	33.1	59.9	71.5	82.2	**68.6**	**89.3**	**93.9**	**97.2**	**84.4**	**98.2**	**99.8**	**100**
WHOS+LOMO	53.0	79.8	89.6	95.5	71.2	90.0	94.4	97.3	83.6	95.4	98.8	**100**
CNN+LOMO	49.9	77.5	86.9	93.8	79.8	93.6	96.3	98.2	88.0	98.3	99.5	**100**
WHOS+CNN	49.7	78.0	87.9	94.4	76.1	92.9	96.1	98.2	**89.0**	98.5	**99.6**	**100**
WHOS+LOMO+CNN	**54.6**	**81.5**	**90.3**	**95.7**	**80.8**	**94.6**	**96.9**	**98.7**	88.4	**98.8**	**99.6**	**100**

Dataset	CUHK03 (Manually)				CUHK03 (Detected)				Market-1501			
Metric (%)	R1	R5	R10	R20	R1	R5	R10	R20	R1(S)	mAP(S)	R1(M)	mAP(M)
WHOS [131]	59.6	87.2	92.8	96.9	50.6	82.1	90.4	96.1	55.2	27.5	60.3	33.5
LOMO [24]	61.6	87.0	94.6	98.0	53.4	83.1	91.2	96.4	57.7	29.0	68.0	37.8
CNN [9]	**81.5**	**95.7**	**97.1**	**98.0**	**80.3**	**96.3**	**98.6**	**99.0**	**72.7**	**48.1**	**80.2**	**58.5**
WHOS+LOMO	65.2	92.2	96.8	**99.1**	59.9	89.4	95.5	98.5	62.4	33.6	69.0	41.0
CNN+LOMO	**82.6**	96.0	97.5	98.6	82.4	95.7	97.4	98.4	73.0	48.5	80.9	59.1
WHOS+CNN	80.4	95.7	98.0	98.4	81.1	95.4	97.5	**98.6**	72.8	48.3	80.3	58.7
WHOS+LOMO+CNN	81.9	**96.5**	**98.2**	98.9	**83.3**	**96.2**	**97.9**	**98.6**	**73.9**	**49.4**	**81.4**	**59.9**

Table 5.7: Effects of feature choice in re-id performance using the IRS model with Uniform Coding.

When a single type of feature is used, it is found that deep CNN feature gives the best re-id performance, except on VIPeR, and LOMO is more discriminative than WHOS most of the time. The advantage of CNN deep feature over both hand-crafted features LOMO and WHOS is very significant given larger training data in CUHK03 and Market-1501, yielding Rank-1 rate *increase* of 19.9% (CUHK03 (Manual)), 26.9% (CUHK03 (Detected)), and 15.0% (Market-1501) against LOMO. Without fine-tuning a CUHK03 trained CNN deep feature on the target domains, it still performs the best on CUHK01 due to the considerable similarity in viewing conditions between CUHK01 and CUHK03. CNN feature performs less well on VIPeR given the greater discrepancy in viewing conditions between VIPeR and CUHK03, similar to the domain shift problem in transfer learning [65, 144].

We further evaluated multi-feature based re-id performance by score-level fusion. It is evident that most combinations lead to improved person re-id performance, and fusing all three features often generate the best accuracies. This observation confirms the previous findings that different appearance information can be encoded by distinct features and their fusion enhances the effect of each other [46, 47, 96, 136].

(IV) Regularisation Sensitivity. We analysed the sensitivity of the only free parameter λ in Eq. (5.3) which controls the regularisation strength of our IRS model. This evaluation was conducted with the LOMO feature and multi-query setting on Market-1501 [2]. Specifically, we evaluated the Rank-1 and mAP with λ varying from 0 to 0.1. Figure 5.6 shows that the performance of our IRS model is not sensitive to λ, with a large satisfactory range.

(V) Model Complexity. In addition to model re-id accuracy, we also examined the model

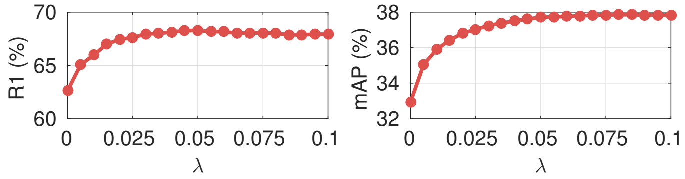


Figure 5.6: Regularisation sensitivity on the Market-1501 dataset [2], with the multi-query setting used.

Dataset	VIPeR	CUHK01	CUHK03	Market-1501
Training Size	632	1940	12197	12936
MLAPG	50.9	746.6	4.0×10^4	-
kLFDA	5.0	45.9	2203.2	1465.8
XQDA	4.1	51.9	3416.0	3233.8
NFST	1.3	6.0	1135.1	801.8
IRS	**1.2**	**4.2**	**248.8**	**266.3**

Table 5.8: Model complexity and training costs of person re-id models. *Metric*: Model training time (in seconds), smaller is better.

complexity and computational costs, in particular model training time. We carried out this evaluation by comparing our IRS model with some strong metric learning methods including kLFDA [23], XQDA [24], MLAPG [25], and NFST [47]. Given n training samples represented by d-dimensional feature vectors, it requires $\frac{3}{2}dnm+\frac{9}{2}m^3$ ($m=\min(d,n)$) floating point addition and multiplications [145] to perform an eigen-decomposition for solving either a generalised eigen-problem [23, 24] or a null space [47], whereas solving the linear system (Eq. (5.3)) of the IRS model takes $\frac{1}{2}dnm+\frac{1}{6}m^3$ [146]. Deep learning models [50, 9, 52] are not explicitly evaluated since they are usually much more demanding in computational overhead, requiring much more training time (days or even weeks) and more powerful harware (GPU). In this evaluation, we adopted the LOMO feature for all datasets and all the models compared, the 485/486 people split on CUHK01, the manually labelled person bounding boxes on CUHK03, and the single-query setting on Market-1501.

For each model, we recorded and compared the average training time of 10 trials performed on a Linux OS based workstation with 2.6GHz CPU. Table 5.8 presents the training time of different models (in seconds). On the smaller VIPeR dataset, our IRS model training needed only 1.2 seconds, similar at NFST and 42.4 times faster than MLAPG. On larger datasets CUHK01, CUHK03 and Market-1501, all models took longer time to train and training the IRS model remains the fastest with speed-up over MLAPG enlarged to 177.8 / 160.8 times on CUHK01

Dataset		VIPeR					CUHK01					CUHK03					Market-1501				
Label #		50	100	150	200	*ALT*	50	100	150	200	*ALT*	50	100	150	200	*ALT*	50	100	150	200	*ALT*
Time (sec.)	BL	0.23	0.23	0.25	0.26	36.5	1.43	1.51	1.57	1.66	232.8	20.4	21.7	22.4	24.5	3349.9	119.5	121.5	125.6	140.3	1.9×10^4
	IL	**0.02**	**0.02**	**0.02**	**0.03**	**3.28**	**0.14**	**0.15**	**0.16**	**0.17**	**23.4**	**1.62**	**1.69**	**1.70**	**1.81**	**257.0**	**1.94**	**5.05**	**6.61**	**9.60**	**877.3**
R1 (%)	BL	**20.6**	**29.2**	**34.9**	**38.9**	-	**21.9**	**37.3**	**46.5**	**52.5**	-	**24.0**	**35.2**	**40.5**	**43.8**	-	**28.6**	**44.5**	**51.7**	**55.2**	-
	IL	19.4	**29.2**	33.6	37.2	-	20.8	35.6	45.3	51.5	-	22.1	33.0	38.8	41.7	-	27.5	44.2	50.6	54.3	-

Table 5.9: Comparing passive Incremental Learning (IL) vs. Batch-wise Learning (BL) using the IRS model. ALT: Accumulated Learning Time, i.e. the summed time for training all the 151 IRS models when the label number is increased from 50 to 200 one by one.

Dataset	VIPeR				CUHK01				CUHK03				Market-1501			
Label #	50	100	150	200	50	100	150	200	50	100	150	200	50	100	150	200
Random	19.4	29.2	33.6	37.2	20.8	35.6	45.3	51.5	22.1	33.0	38.8	41.7	27.5	44.2	50.6	54.3
Density [81]	18.4	26.8	33.5	37.5	23.3	37.0	44.5	50.0	23.7	34.8	40.2	42.7	32.3	46.2	51.5	53.9
JointE^2	**23.4**	**31.4**	**36.5**	**40.9**	**29.9**	**39.7**	**47.1**	**52.2**	**25.1**	**36.8**	**41.3**	**43.0**	**36.5**	**50.7**	**54.8**	**58.2**

Table 5.10: Evaluation on the active incremental learning algorithm.

/ CUHK03, respectively[2]. This demonstrates the advantage of the proposed IRS model over existing competitors for scaling up to large sized training data.

5.6.2 Incremental Person Re-Identification Evaluation

We further evaluated the performance of our IRS model using the incremental learning IRS$^{\text{inc}}$ algorithm (Section 5.3). This setting starts with a small number, e.g. 10 of labelled true match training pairs, rather than assuming a large pre-collected training set. More labelled data will arrive one by one over time during deployment due to human-in-the-loop verification. In such a setting, a re-id model can naturally evolve through deployment life-cycle and efficiently adapt to each application test domain. In this context, we consider two incremental re-id model learning scenarios: **(I)** *Passive* incremental learning where unlabelled person images are randomly selected for human to verify; **(II)** *Active* incremental learning where person images are actively determined by the proposed JointE^2 active learning algorithm as detailed in Section 5.4.

Incremental Re-Id Evaluation Protocol. Due to the lack of access to large sized training samples in batch, incrementally learned models are typically less powerful than batch learned models [87, 92]. Therefore, it is critical to evaluate how much performance drop is introduced by the Incremental Learning (IL) algorithm, IRS$^{\text{inc}}$, as compared to the corresponding Batch-wise Learning (BL) and how much efficiency is gained by IL. We started with 10 labelled identities, i.e. cross-camera truth matches of 10 persons, and set the total labelling budget to 200 persons. For simplicity, we selected four test cases with $50, 100, 150, 200$ labelled identities respectively and

[2] The MLAPG model failed to converge on Market-1501.

evaluated their model accuracy and training cost. To compare the Accumulated Learning Time (ALT), i.e. the summed time for training all the IRS models when the label number is increased from 50 to 200 one by one (in total 151 updates), we interpolated estimations on training time between these four measured test cases. We adopted the LOMO feature on all datasets. We utilised the 485/486 people split on CUHK01, the manually labelled person images on CUHK03, the single-query setting on Market-1501, and the same test data as the experiments in Section 5.6.1. We conducted 10 folds of evaluations each with a different set of random unlabelled identities and reported the averaged results.

(I) Passive Incremental Learning. We compared the proposed incremental learning (IL) based IRS (IRS^{inc}) with batch-wise learning (BL) based IRS in Table 5.9 for model training time and re-id Rank-1 performance. It is found that IRS model training speed can increase by one order of magnitude or more, with higher speed-up observed on larger datasets and resulting in more model training efficiency gain. Specifically, on VIPeR, BL took approximately 36.5 seconds to conduct the 151 model updates by re-training, whereas IL only required 3.28 seconds. When evaluated on Market-1501, BL took over 5.5 hours (1.9×10^4 seconds) to perform the sequential model updates, while IL was more than 20× faster, only took 877.3 seconds. Importantly, this speed-up is at the cost of only $1 \sim 2\%$ Rank-1 drop. This suggests an attractive trade-off for the IRS^{inc} algorithm between effectiveness and efficiency in incremental model learning.

(II) Active Incremental Learning. We further evaluated the effect of the proposed *JointE*^{2} active learning algorithm (Section 5.4) by random passive unlabelled image selection (*Random*). Also, we compared with a state-of-the-art density based active sampling method [81] which prefers to query the densest region of unlabelled sample space (*Density*). For both active sampling methods, we used our IRS^{inc} for re-id model training. We evaluated the four test cases as shown in Table 5.9.

It is evident from Table 5.10 that: **(1)** On all four datasets, our JointE^2 outperformed clearly both *Random* and *Density* given varying numbers of labelled samples. For example, when 50 identities were labelled, the proposed JointE^2 algorithm beats *Random* sampling in Rank-1 by 4.0%(23.4−19.4), 9.1%(29.9−20.8), 3.0%(25.1−22.1), 9.0%(36.5−27.5) on VIPeR, CUHK01, CUHK03 and Market-1501, respectively. **(2)** Our JointE^2 model obtained similar or even better performance with less human labelling effort. For example, on Market-1501, by labelling 150 identities, JointE^2 achieved Rank-1 rate of 54.8%, surpassed Random (54.3%) and Density

(53.9%) with a greater budget of 200 identities.

In summary, the results in Tables 5.9 and 5.10 show clearly that the hybrid of our proposed IRS$^{\text{inc}}$ model and JointE^2 active sampling method provides a highly scalable active incremental re-id model training framework, with attractive model learning capability and efficiency from less labelling effort suited for real-world person re-id applications.

5.7 Summary

In this chapter, we developed a novel approach to explicitly designing a feature embedding space for supervised person re-identification model optimisation. We solved the re-id model learning problem by introducing an identity regression method in an Identity Regression Space (IRS) with an efficient closed-form solution. Furthermore, we formulated an incremental learning algorithm IRS$^{\text{inc}}$ to explore sequential on-line labelling and model updating. This enables the model to not only update efficiently the re-id model once new data annotations become available, but also improve adaptively the re-id model to new test domains. To better leverage human annotation effort, we further derived a novel active learning method JointE^2 to selectively query the most informative unlabelled data online. Extensive experiments on four benchmarks show that our IRS method outperforms existing state-of-the-art re-id methods in the conventional batch-wise model learning setting. Moreover, the proposed incremental learning algorithm increases significantly model training speed, over 10 times faster than batch-wise model learning, by only sacrificing marginal model re-id capability with $1 \sim 2\%$ Rank-1 drop. Our active learning method improves notably the human labelling quality, particularly when limited budget is accessible, providing over 3% Rank-1 improvement than Random sampling given 50 identities labelling budget.

Chapter 6

Human-In-The-Loop Learning from Relevance Feedback

6.1 Overview

The previous chapters have investigated techniques to minimise human labelling efforts spent during the model training stage, i.e. either by unsupervised learning directly from unlabelled data (Chapter 3, 4), or by active learning on a small group of representative data labelled by human (Chapter 5). However, they ignored one important aspect of human labelling in a re-id system, which is that human efforts could also be required in the model deployment stage. In particular, in real-world scenarios where the population size in the potential searching space is very large, even the current best re-id model still cannot achieve satisfiable performance for fully-automated deployments. We observed on CUHK03 dataset that (Section 6.5), a 10-fold increase in gallery size leads to a 10-fold decrease in re-id Rank-1 performance (i.e. single-digit Rank-1 accuracy). Given such low Rank-1 scores, in practice human operators (users) are still required to verify any true match of a probe from the output ranking list generated by any re-id model.

In this chapter, we aim to save such human efforts spent in the deployment stage, by formulating a hybrid human-computer learning paradigm with humans in the model matching loop (Fig. 6.1(c)). We call this semi-automated scheme *Human-In-the-Loop* (HIL) re-id, designed to optimise re-id performance on a larger-sized test population (either with or without training data), as compared to the conventional *Human-Out-of-the-Loop* (HOL) re-id models that are mostly designed to optimise re-id given a larger size labelled training data and a small size test population. This HIL re-id scheme has three significant advantages over the conventional HOL models:

1. *Less human labelling effort*: HIL re-id requires much less human labelling effort, since it does not necessarily require the expensive construction of a pre-labelled training set. More importantly, it prioritises directly the human labour effort on each given re-id task in deployment, rather than optimising the model learning error on an independent training set. More specifically, the number of feedback from human verification is typically in *tens* as compared to *thousands* of offline pre-labelled training data required by HOL methods.

2. *Model transfer learning*: Our HIL model is able to achieve greater transferability with better re-id performance in test domains. This is because a HIL model focuses on re-id matching optimisation directly in the deployment gallery population, rather than learning a distance metric from a separate training set and *assuming* its blind transferability to independent (unseen) test data. It enables a human operator to interactively validate model matching results for each re-id task and inform on model mistakes (similar in spirit to negative mining).

3. *Reinforcing visual consistency*: As computer vision algorithms are intrinsically very different from the human visual system, a re-id model can make mistakes that generate "unexpected" (visually inconsistent) re-id ranking results, readily identifiable by a human observer. By learning directly from the inconsistency between a computer vision model and human observation, a HIL re-id model is guided to maximise visually more consistent ranking lists favoured by human observations, and thus more effective to users of a re-id system.

The main **contribution** of this chapter is a novel HIL re-id model that enables a user to re-identify rapidly a given probe person image after only a handful of feedback verifications even when the search gallery size is large. More specifically, a *Human Verification Incremental Learning* (HVIL) model (Fig. 6.1(c)) is formulated to *simultaneously* minimise human-in-the-loop feedback and maximise model re-id accuracy by incorporating:

1. *Sparse feedback* - HVIL allows for easier human feedback on a few dissimilar matching results without the need for exhaustive eyeball search of true/false in the entire rank list. It aims to rectify rapidly model mistakes by focusing *only* on minimising visually obvious errors (hard negatives) identified by human observation. This is reminiscent to learning by hard negative mining but *with* human in the loop, so to improve model learning with less

training data.

2. *Immediate benefit* - HVIL introduces a new online incremental distance metric learning algorithm, which enables real-time model response to human feedback by rapidly presenting a freshly optimised ranking list for further human feedback, quickly leading to identifying a true match.

3. *The older the wiser* - HVIL is updated cumulatively on-the-fly utilising multiple user feedback per probe, with incremental model optimisation for each new probe given what have been learned from all previous probes.

4. *A strong ensemble model* - An additional Regularised Metric Ensemble Learning (RMEL) model is introduced by taking all the incrementally optimised per-probe models as a set of "weak" models [147, 148] and constructing a "strong" ensemble model for performing HOL re-id tasks when human feedback becomes unavailable.

Extensive comparative experiments on three benchmark datasets (CUHK03 [1], Market-1501 [2], and VIPeR [7]) demonstrate that this HVIL model outperforms the state-of-the-art methods for both the proposed new HIL and the conventional HOL re-id deployments.

6.2 Human-In-the-Loop Incremental Learning

6.2.1 Problem Definition

Let a person image be denoted by a feature vector $\boldsymbol{x} \in \mathbb{R}^d$. The *Human-In-the-Loop (HIL) re-id* problem is formulated as:

1. For each image $\boldsymbol{x}^p$ in a probe set $\mathcal{P} = \{\boldsymbol{x}_i^p\}_{i=1}^{N_p}$ (Fig. 6.3(a)), $\boldsymbol{x}^p$ is matched against a gallery set $\mathcal{G} = \{\boldsymbol{x}_i^g\}_{i=1}^{N_g}$ and an initial ranking list for all gallery images is generated by a re-id ranking function $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$, according to ranking scores $f_{\boldsymbol{x}^p}(\boldsymbol{x}_i^g)$ (Fig. 6.3(b)).

2. A human operator (user) browses the gallery ranking list to verify the existence and the rank of any true match for $\boldsymbol{x}^p$. Human feedback is generated when a ranked gallery image $\boldsymbol{x}^g$ is selected by the user with a label $y \in \{\text{true}, \text{dissimilar}\}$ (Fig. 6.3(c)). Once a feedback on probe $\boldsymbol{x}^p$ is received, the parameters of re-id model $f(\cdot)$ are updated instantly (Fig. 6.3(d)) to re-order the gallery ranking list and give the user immediate reward for the next feedback (Fig. 6.3(e)).

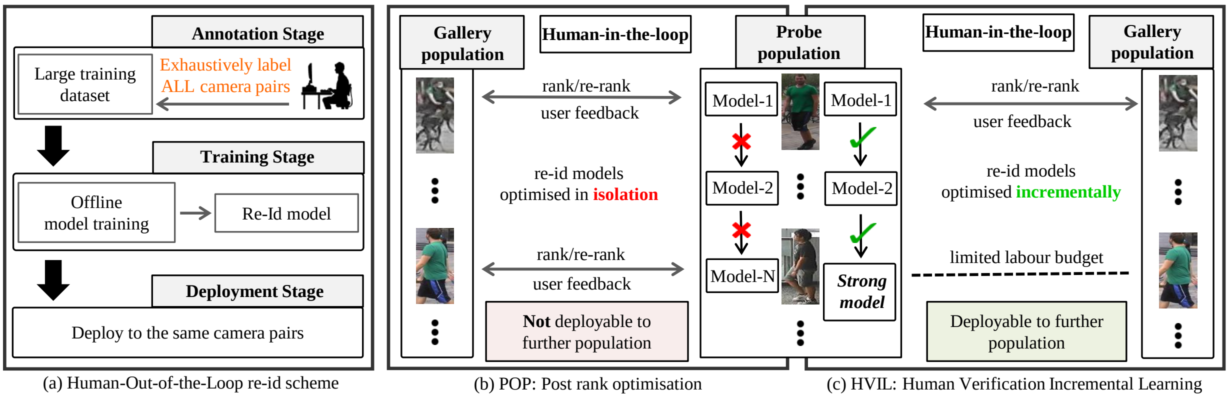


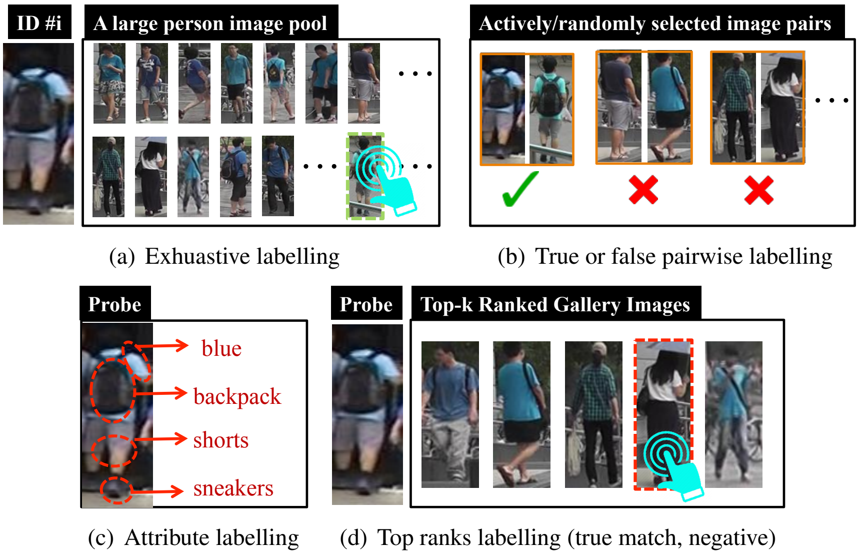
Figure 6.1: Illustration of two person re-id schemes. **(a)** The conventional *Human-Out-of-the-Loop* (HOL) re-id scheme requires exhaustive pre-labelled training data collection for supervised model learning. The learned model is assumed sufficiently accurate and then deployed to perform fully automated person re-id tasks without human in the loop. **(b)** POP [69]: A recent *Human-In-the-Loop* (HIL) re-id approach which optimises probe-specific models in isolation from human feedback verifications in the deployment time. All probe people requires human in the loop. **(c)** HVIL: The proposed new incremental HIL re-id model capable of not only progressively learning a generalised model from human verifications across all probed people while carrying out the HIL re-id tasks, but also performing the HOL re-id tasks when human effort becomes unavailable.

3. When either a true match is found or a pre-determined maximum round of feedback is reached, the next probe is presented for re-id matching in the gallery set. In contrast to pre-labelling training data required by the conventional train-once-and-deploy *human-out-of-the-loop (HOL) re-id* scheme, *HIL re-id* has two unique characteristics: (a) Due to limited human patience and labour budget [107], a user typically prefers to examine only the top ranks rather than the whole rank list, and to provide only a few feedback. (b) Instead of seeking to verify *positives* (true matches) for each probe, which are most *unlikely* to appear in the top ranks[1], it is a much easier and more rewarding task for the user to identify *strong-negatives*, that is, those top ranked negative gallery instances *“definitely not the one I am looking for”* – visually very *dissimilar* to the target image.

Note that, in contrast to [69, 48], here we consider a simpler human verification task by also ignoring *weak-negatives*: Those top ranked negative instances which *“look similar but not the same person as I am looking for”*. The reasons are:

1. A user is inclined to notice strong negatives among the top ranks, i.e. a cognitively easier task (Fig. 6.2(d)) due to that most top ranks are likely to be weak negatives. Making

[1] In a large size gallery set, true matches are often scarce (only one-shot or few shots) and overwhelmed (appear in low-ranks) by false matches of high-ranks in the rank list.


ID #i
A large person image pool
Actively/randomly selected image pairs
✓
✗
✗
Probe
blue
backpack
shorts
sneakers
Probe
Top-k Ranked Gallery Images

(a) Exhuastive labelling (b) True or false pairwise labelling

(c) Attribute labelling (d) Top ranks labelling (true match, negative)

Figure 6.2: Different human labelling processes are employed in person re-id model training and deployment. **(a)** Large size offline labelling of cross-view positive- and negative-pairs of training data with identity labels [46, 25, 95, 47]. **(b)** Selective or random sampling of person image pairs for human verification in either model training [86] or deployment [85]. **(c)** Fine-grained attribute labelling in either training [143] or deployment [85]. **(d)** True match verification among the top ranked sub-list in model deployment [107, 69, 48], or verification of both visually dissimilar and similar wrong matches in top ranks (strong/hard and weak negative mining) in model deployment [69, 48].

correct selection and verification of all weak negatives requires much more effort. In contrast, a strong negative "stands out" readily to a user's attention among the top ranks given the salience-driven visual selective attention mechanism built into the human visual system [149].

2. We consider strong negatives in top ranks are *hard-unexpected negatives*: "Hard" since they are top-ranked negatives in the gallery thus misclassifed with high confidence (short matching distance) to the wrong identity class by the current model; "Unexpected" since they are visually significantly dissimilar to the probe image whilst among the top ranks, therefore violating expectation and providing most informative feedback on model mistakes[2]. Exploiting strong negatives to rectify model learning is more cost-effective with less labelling required (Section 6.5). Moreover, this is also compatible with the notion of salience-guided human eye movements therefore more likely to encourage a user to engage with the re-id task at hand whilst giving feedback, providing a higher degree of complementary effect between iterative machine learning from human feedback and hu-

[2] In this context, weak negatives in top ranks can be considered as *hard-expected negatives* [150].

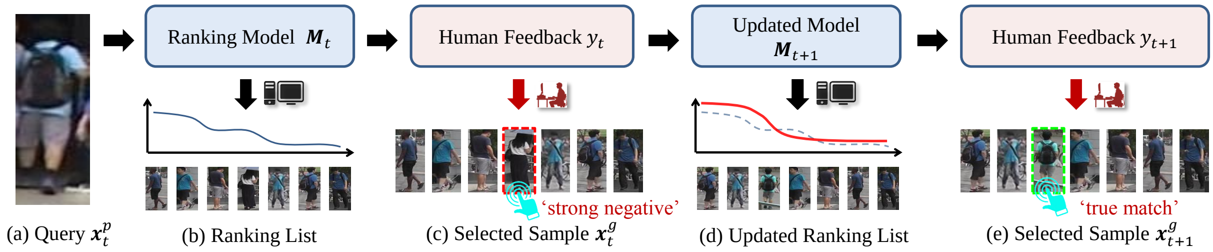

Ranking Model $\boldsymbol{M}_t$
Human Feedback y_t
Updated Model $\boldsymbol{M}_{t+1}$
Human Feedback y_{t+1}
'strong negative'
'true match'
(a) Query $\boldsymbol{x}_t^p$
(b) Ranking List
(c) Selected Sample $\boldsymbol{x}_t^g$
(d) Updated Ranking List
(e) Selected Sample $\boldsymbol{x}_{t+1}^g$

Figure 6.3: Visualisation of the proposed Human-In-the-Loop person re-id procedure.

man rewards from improved model output.

6.2.2 Modelling Human Feedback

Formally, we wish to construct an incrementally optimised ranking function, $f_{\boldsymbol{x}^p}(\boldsymbol{x}_i^g) : \mathbb{R}^d \rightarrow \mathbb{R}$, where $f(\cdot)$ can be estimated by two types of human feedback $y \in L = \{m, s\}$ as *true-match* and *strong-negative* respectively. Inspired by [151, 152, 153], we define a ranking error (loss) function for a feedback y on a human selected gallery sample $\boldsymbol{x}^g$ given a probe $\boldsymbol{x}^p$ as:

$$err(f_{\boldsymbol{x}^p}(\boldsymbol{x}^g), y) = \mathcal{L}_y(rank(f_{\boldsymbol{x}^p}(\boldsymbol{x}^g))), \tag{6.1}$$

where $rank(f_{\boldsymbol{x}^p}(\boldsymbol{x}^g))$ denotes the rank of $\boldsymbol{x}^g$ given by $f_{\boldsymbol{x}^p}(\cdot)$, defined as:

$$rank(f_{\boldsymbol{x}^p}(\boldsymbol{x}^g)) = \sum_{\boldsymbol{x}_i^g \in \boldsymbol{G} \setminus \boldsymbol{x}^g} \mathcal{I}(f_{\boldsymbol{x}^p}(\boldsymbol{x}_i^g) \geqslant f_{\boldsymbol{x}^p}(\boldsymbol{x}^g)), \tag{6.2}$$

where $\mathcal{I}(\cdot)$ is the indicator function. The loss function $\mathcal{L}_y(\cdot) : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ transforms a rank into a loss. We introduce a novel re-id ranking loss defined as:

$$\mathcal{L}_y(k) = \begin{cases} \sum_{i=1}^{k} \alpha_i, & \text{if } y \in \{m\} \\ \sum_{i=k+1}^{n_g} \hat{\alpha}_i, & \text{if } y \in \{s\} \end{cases}, \tag{6.3}$$
$$\text{with } \alpha_1 \geqslant \alpha_2 \geqslant \cdots \geqslant 0, \text{ and } \hat{\alpha}_{n_g} \geqslant \hat{\alpha}_{n_g-1} \geqslant \cdots \geqslant 0.$$

Note, different choices of $\alpha_i, \hat{\alpha}_i$ lead to specific model responses to human feedback (Fig. 6.4). We set $\alpha_i = \frac{1}{i}$ (large penalty with steep slope) when y indicates a *true-match* (m), and $\hat{\alpha}_i = \frac{1}{n_g - 1}$ with n_g the gallery size (small penalty with gentle slope) when y represents a *strong-negative* (s). Such a ranking loss is designed to favour a model update behaviour so that: (1) *true-matches* are quickly pushed up to the top ranks, whilst (2) *strong-negatives* are mildly moved towards the bottom rank direction. Our experiments (Sec. 6.5.1) show that such a ranking loss criterion boosts very effectively the Rank-1 matching rate and pushes quickly *true-matches* to the top ranks at each iteration of human feedback.

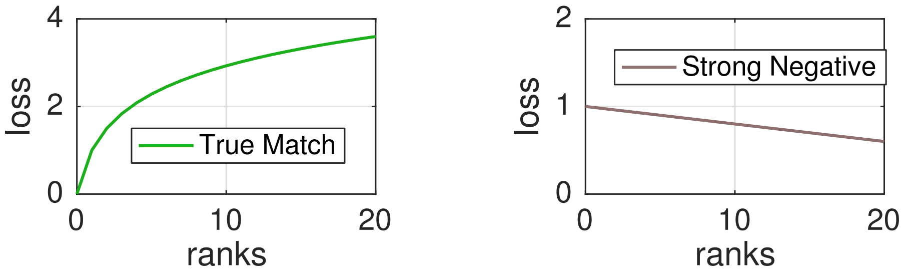


Figure 6.4: Values of $\mathcal{L}_y(k)$ for distinct human feedback, with $n_g = 50$.

6.2.3 Real-Time Model Update

Model Formulation Given the re-id ranking loss function defined in Eqn. (6.3), we wish to have real-time model update to human feedback therefore providing instant reward to user labour effort. To that end, we consider the HVIL re-id ranking model $f(\cdot)$ as a negative Mahalanobis distance metric:

$$f_{\boldsymbol{x}^p}(\boldsymbol{x}^g) = -\left[(\boldsymbol{x}^p - \boldsymbol{x}^g)^\top \boldsymbol{M}(\boldsymbol{x}^p - \boldsymbol{x}^g)\right], \ \boldsymbol{M} \in S^d_+. \tag{6.4}$$

The positive semi-definite matrix $\boldsymbol{M}$ consists of model parameters to be learned.

Knowledge Cumulation by Online Learning In the previous works [69, 107], a re-id model $f(\cdot)$ is only optimised in isolation for each probe without benefiting from previous feedback on other probes. To overcome this limitation, we wish to optimise $f(\cdot)$ *incrementally* in an online manner [154] for maximising the value of limited human feedback labour budget. Moreover, to achieve real-time human-in-the-loop feedback and reward, $f(\cdot)$ needs be estimated immediately on each human feedback.

Formally, given a new probe $\boldsymbol{x}^p_t$ at time step $t \in \{1, \cdots, \tau\}$ (τ the pre-defined verification budget), a user is presented with a gallery rank list computed by the previously estimated model $\boldsymbol{M}_{t-1}$ instead of a new ranking function re-initialised from scratch for this new probe. The user then verifies a gallery image $\boldsymbol{x}^g_t$ in the top ranks with a label y_t, generating a labelled triplet $(\boldsymbol{x}^p_t, \boldsymbol{x}^g_t, y_t)$. Given Eqn. (6.3), this triplet has a corresponding loss as $\mathcal{L}^{(t)} = \mathcal{L}_{y_t}(rank(f_{\boldsymbol{x}^p_t}(\boldsymbol{x}^g_t)))$.

We update the ranking model by minimising the following objective function:

$$\boldsymbol{M}_t = \underset{\boldsymbol{M} \in S_+^d}{\operatorname{argmin}} \Delta_F(\boldsymbol{M}, \boldsymbol{M}_{t-1}) + \eta \mathcal{L}^{(t)}, \tag{6.5}$$

where Δ_F is a Bregman divergence measure, defined by an arbitrary differentiable convex function F, for quantifying the discrepancy between $\boldsymbol{M}$ and $\boldsymbol{M}_{t-1}$. The set S_+^d defines a positive semi-definite (PSD) cone. The tradeoff parameter $\eta > 0$ balances the model update divergence and empirical ranking loss. This optimisation updates the ranking model adopted from the previous probe by encoding user feedback on the current probe.

Loss Approximation for Real-Time Optimisation In order to encourage and maintain user engagement in verification feedback, real-time online incremental metric learning is required. However, as $\mathcal{L}^{(t)}$ is discontinuous, the overall objective function cannot be optimised efficiently by gradient-based learning methods. We thus approximate the loss function by a continuous upper bound [151] so that it is differentiable w.r.t. the parameter $\boldsymbol{M}$:

$$\begin{aligned}\widetilde{\mathcal{L}}^{(t)} = \frac{1}{\mathcal{N}_t^-} \sum_{\boldsymbol{x}_i^g \in \boldsymbol{G} \setminus \boldsymbol{x}_t^g} & \mathcal{L}_{y_t}\left(rank\left(f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g | \boldsymbol{M}_{t-1})\right)\right) \\ & \cdot h_{y_t}\left(f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g | \boldsymbol{M}_t) - f_{\boldsymbol{x}^p}(\boldsymbol{x}_i^g | \boldsymbol{M}_{t-1})\right)^2,\end{aligned} \tag{6.6}$$

where $f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g | \boldsymbol{M}_{t-1})$ denotes the function value of $f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g)$ parametrised by $\boldsymbol{M}_{t-1}$, and $h_{y_t}(\cdot)$ represents a hinge loss function defined as:

$$\begin{aligned} & h_{y_t}\left(f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g) - f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_i^g)\right) = \\ & \begin{cases} max\big(0, 1 - f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g) + f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_i^g)\big), & \text{if } y_t \in \{m\} \\ max\big(0, 1 - f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_i^g) + f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g)\big), & \text{if } y_t \in \{s\} \end{cases} . \end{aligned} \tag{6.7}$$

The normaliser $\mathcal{N}_t^-$ in Eqn. (6.6) is the amount of violators, i.e. the gallery instances that generate non-zero hinge loss (Eqn. (6.7)) w.r.t. the triplet $(\boldsymbol{x}_t^p, \boldsymbol{x}_t^g, y_t)$.

Learning Speed-up by Most Violator Update Given the loss approximation in Eqn. (6.6), we can exploit the stochastic gradient descent (SGD) algorithm [155] for optimising the proposed model objective function Eqn. (6.5) by iteratively updating on sub-sampled batches of all violators. However, the computational overhead of iterative updates can be high due to possibly large number of violators, and thus not meeting the real-time requirement. To address this problem, we explore a *most violator update* strategy, that is, to perform metric updates using *only* the violator $\boldsymbol{x}_v^g$ with the most violation (Eqn. (6.7)). The final approximated empirical loss is then estimated as:

$$\begin{aligned}\widetilde{\mathcal{L}}_v^{(t)} = \mathcal{L}_{y_t} \Big(rank\big(f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_{t-1})\big)\Big) \\ \cdot h_{y_t} \Big(f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_t) - f_{\boldsymbol{x}^p}(\boldsymbol{x}_v^g|\boldsymbol{M}_{t-1})\Big)^2 .\end{aligned} \tag{6.8}$$

Next, we derive $\boldsymbol{M}_t$ for updating the ranking metric. Specifically, recall that the Bregman divergence between any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined as:

$$\Delta_F(\boldsymbol{A},\boldsymbol{B}) = F(\boldsymbol{A}) - F(\boldsymbol{B}) - tr\big((\boldsymbol{A}-\boldsymbol{B})g(\boldsymbol{B})^\top\big), \tag{6.9}$$

where $g(\cdot)$ denotes the derivative of F (Eqn. (6.5)) [156] and $tr(\cdot)$ the matrix trace norm. After taking the gradient with the first argument $\boldsymbol{A}$, it has the following form:

$$\nabla_{\boldsymbol{A}}\Delta_F(\boldsymbol{A},\boldsymbol{B}) = g(\boldsymbol{A}) - g(\boldsymbol{B}), \tag{6.10}$$

By replacing $\mathcal{L}^{(t)}$ in Eqn. (6.5) with $\widetilde{\mathcal{L}}_v^{(t)}$, and setting the gradient of the minimisation objective in Eqn. (6.5) to zero, we have:

$$g(\boldsymbol{M}_t) - g(\boldsymbol{M}_{t-1}) + \eta\nabla_{\boldsymbol{M}}\widetilde{\mathcal{L}}_v^{(t)} = 0. \tag{6.11}$$

This gives the following ranking metric online updating criterion:

$$\boldsymbol{M}_t = g^{-1}\Big(g(\boldsymbol{M}_{t-1}) - \eta\nabla_{\boldsymbol{M}}\widetilde{\mathcal{L}}_v^{(t)}\Big), \tag{6.12}$$

where the gradient of $\widetilde{\mathcal{L}}_v^{(t)}$ w.r.t. $\boldsymbol{M}$ can be calculated as:

$$\nabla_{\boldsymbol{M}}\widetilde{\mathcal{L}}_v^{(t)} = \hat{\mathcal{L}}(f_t - f_v - b_t)\boldsymbol{z}_t\boldsymbol{z}_t^\top, \tag{6.13}$$

with

$$\hat{\mathcal{L}} = \mathcal{L}_{y_t}\left(rank\big(f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_{t-1})\big)\right),\ f_v = f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_v^g|\boldsymbol{M}_{t-1}), \tag{6.14}$$

$$f_t = f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_t),\ \ \boldsymbol{z}_t = \boldsymbol{x}_t^p - \boldsymbol{x}_t^g,\ b_t = \begin{cases} 1, & \text{if } y_t \in \{m\}. \\ -1, & \text{if } y_t \in \{s\}. \end{cases}$$

For the convex function $F(\cdot)$, existing common choices include squared Frobenius norm $\|\boldsymbol{M}\|_F^2$ and quantum entropy $tr(\boldsymbol{M}\log(\boldsymbol{M}) - \boldsymbol{M})$. The incremental update of the HVIL model by Eqn. (6.12) can be optimised by a standard gradient-based learning scheme such as [151, 157, 156]. In this work, we adopt a strictly convex function $F(\boldsymbol{M}) = -\log\det(\boldsymbol{M})$. This is because its gradient function $g(\cdot)$ is as simple as

$$g(\boldsymbol{M}) = \nabla_{\boldsymbol{M}} F(\boldsymbol{M}) = \boldsymbol{M}^{-1}, \tag{6.15}$$

and along with Eqn. (6.13) we can simplify Eqn. (6.12) as:

$$\boldsymbol{M}_t = \left(\boldsymbol{M}_{t-1}^{-1} - \eta\hat{\mathcal{L}}(f_t - f_v - b_t)\boldsymbol{z}_t\boldsymbol{z}_t^\top\right)^{-1}. \tag{6.16}$$

Applying the Sherman Morrison formula [125], we obtain the following online updating scheme for our HVIL model $\boldsymbol{M}$:

$$\boldsymbol{M}_t = \boldsymbol{M}_{t-1} - \frac{\eta\hat{\mathcal{L}}(f_t - f_v - b_t)\boldsymbol{M}_{t-1}\boldsymbol{z}_t\boldsymbol{z}_t^\top\boldsymbol{M}_{t-1}}{1 + \eta\hat{\mathcal{L}}(f_t - f_v - b_t)\boldsymbol{z}_t^\top\boldsymbol{M}_{t-1}\boldsymbol{z}_t} \tag{6.17}$$

To compute $\boldsymbol{M}_t$, we need to obtain the value of f_t which however is parametrised by $\boldsymbol{M}_t$ (Eqn. (6.14)) and thus cannot be computed readily. One potential optimisation option is resorting to gradient approximation [158]. Instead, we propose to solve $\boldsymbol{M}_t$ with exact gradient for more accurate modelling, inspired by the LEGO metric update [159]. Specifically, by left multiplying $\boldsymbol{M}_t$ with $\boldsymbol{z}^\top$ and right multiplying with $\boldsymbol{z}$, we obtain

$$\boldsymbol{z}^\top\boldsymbol{M}_t\boldsymbol{z} = f_t = \frac{\hat{f}}{1 + \eta\hat{\mathcal{L}}(f_t - f_v - b_t)\hat{f}} \tag{6.18}$$

with $\hat{f} = f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_{t-1})$. Then, f_t can be solved by algebra transformation as:

$$f_t = \frac{\eta\hat{\mathcal{L}}(f_v + b_t)\hat{f} - 1 + \sqrt{(\eta\hat{\mathcal{L}}(f_v + b_t)\hat{f} - 1)^2 + 4\eta\hat{\mathcal{L}}\hat{f}^2}}{2\eta\hat{\mathcal{L}}\hat{f}} \tag{6.19}$$

Given this explicitly calculated f_t, we can evaluate quantitatively Eqn. (6.17) for online HVIL model updating. An overview of the HVIL online learning process is given in Algorithm 2. The updating scheme as described herein is favourable because it requires no computationally expensive eigen-decomposition to project the updated metric back to the PSD cone, and the positive definiteness of $\boldsymbol{M}_t$ can be automatically guaranteed according to:

Theorem 1. If $\boldsymbol{M}_{t-1}$ is positive definite, then $\boldsymbol{M}_t$ computed by Eqn. (6.17) is also positive definite.

Proof. If $\boldsymbol{M}_{t-1}$ is a positive definite matrix, then

$$\hat{f} = f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_{t-1}) = \boldsymbol{z}_t^\top \boldsymbol{M}_{t-1}\boldsymbol{z}_t > 0 \text{ for all } \boldsymbol{z}_t.$$

Since $\eta > 0$, $\hat{\mathcal{L}} > 0$, we have

$$\sqrt{(\eta\hat{\mathcal{L}}(f_v+b_t)\hat{f}-1)^2+4\eta\hat{\mathcal{L}}\hat{f}^2} > |\eta\hat{\mathcal{L}}(f_v+b_t)\hat{f}-1|.$$

Therefore, from Eqn. (6.19) we have

$$f_t = f_{\boldsymbol{x}_t^p}(\boldsymbol{x}_t^g|\boldsymbol{M}_t) = \boldsymbol{z}_t^\top \boldsymbol{M}_t\boldsymbol{z}_t > 0 \text{ for all } \boldsymbol{z}_t.$$

Hence $\boldsymbol{M}_t$ is also a positive definite matrix.

Model Complexity This online HVIL model update by Eqn. (6.17) is solved with a computational complexity of $\mathcal{O}(d^2)$ where d is the feature vector dimension, while a cost of $\mathcal{O}(d^3)$ is required by most other schemes which perform the Bregman projection back to the PSD cone [151, 157, 156]. Given all the components described above, our final model for Human Verification Incremental Learning (HVIL) enables real-time incremental model learning with human-in-the-loop feedback to model re-id rank list. As shown in our evaluation (Sec. 6.5.1), the proposed HVIL model provides faster human-in-the-loop feedback-reward cycles as compared to alternative models.

6.3 Metric Ensemble for Human-Out-of-the-Loop Re-Identification

Finally, we consider the situation when the limited human labour budget is exhausted at time τ and an automated HOL re-id strategy is required for any further probes as in conventional

Algorithm 2: Human Verification Incremental Learning (HVIL)

```
Input: Unlabelled probe set $\mathcal{P}$ and gallery set $\mathcal{G}$;
Output: Per probe optimised ranking lists; re-id models $\{\boldsymbol{M}_t\}_{t=1}^{\tau}$;
Initialisation: $\boldsymbol{M}_0 = \boldsymbol{I}$ (identity matrix, equivalent to the $L_2$ distance)
HIL person re-id:
while $t < \tau$ do
    Present the next probe $\boldsymbol{x}_t^p \in \mathcal{P}$;
    // maxIter: maximum interactions per probe
    for $iter = 1 : maxIter$ do
        Rank $\mathcal{G}$ with $\boldsymbol{M}_{t-1}$ against the probe $\boldsymbol{x}_t^p$ (Eqn. (6.4));
        Request the human feedback $(\boldsymbol{x}_t^g, y_t)$;
        Calculate $\widetilde{\mathcal{L}}_v^{(t)}$ with the most violator $\boldsymbol{x}_v^g$ (Eqn. (6.7) and (6.8));
        $\boldsymbol{M}_t = update(\boldsymbol{M}_{t-1}, \widetilde{\mathcal{L}}_v^{(t)})$ (Eqn. (6.12));
Return $\{\boldsymbol{M}_t\}_{t=1}^{\tau}$.
```

approaches. In this setting, given that the HVIL re-id model is optimised incrementally during the HIL re-id procedure, the latest model $\boldsymbol{M}_\tau$ optimised by the human verified probe at time τ can be directly deployed. However, it is desirable to construct an even "stronger" model based on metric ensemble learning. Specifically, a side-product of HVIL is a series of models incrementally optimised *locally* for a set of probes with human feedback. We consider them as a set of *globally* "weak" models $\{\boldsymbol{M}_j\}_{j=1}^{\tau}$, and wish to construct a *single globally strong model* for re-identifying further probes without human feedback.

Regularised Metric Ensemble Learning Given weak models $\{\boldsymbol{M}_j\}_{j=1}^{\tau}$, we compute a distance vector $\boldsymbol{d}_{ij} \in \mathbb{R}^\tau$ for any probe-gallery pair $(\boldsymbol{x}_j^g, \boldsymbol{x}_i^p)$:

$$\boldsymbol{d}_{ij} = -\left[f_{\boldsymbol{x}_i^p}(\boldsymbol{x}_j^g|\boldsymbol{M}_1), \cdots, f_{\boldsymbol{x}_i^p}(\boldsymbol{x}_j^g|\boldsymbol{M}_\tau)\right]^\top \tag{6.20}$$

The objective of metric ensemble learning is to obtain an optimal combination of these distances for producing a single globally optimal distance. Here we consider the ensemble ranking function $f_{\boldsymbol{x}_i^p}^{ens}(\boldsymbol{x}_j^g)$ in a bi-linear form (shortened as f_{ij}^{ens}):

$$f_{ij}^{ens} = f_{\boldsymbol{x}_i^p}^{ens}(\boldsymbol{x}_j^g) = -\boldsymbol{d}_{ij}^\top \boldsymbol{W} \boldsymbol{d}_{ij}, \quad \text{s.t.} \quad \boldsymbol{W} \in S_+^\tau, \tag{6.21}$$

with $\boldsymbol{W}$ being the ensemble model parameter matrix that captures the correlations among all the weak model metrics. In this context, previous work such as [46] is a special case of our model when $\boldsymbol{W}$ is restricted to be diagonal only.

Objective Function To estimate an optimal ensemble weights $\boldsymbol{W}$ with maximised identity-discriminative power, we re-use the true matching pairs verified during the human verification procedure (Sec. 6.2) as “training data”: $\mathcal{X}_{tr} = \{(\boldsymbol{x}_i^p, \boldsymbol{x}_i^g)\}_{i=1}^{N_l}$, and their corresponding person identities are denoted by $\mathcal{C} = \{c_i\}_{i=1}^{N_l}$. Note, “training data” here are only for estimating the ensemble model weight, not for learning a distance metric. Since the ranking score f_{ij}^{ens} in Eqn. (6.21) is either negative or zero, we consider that in the extreme case, an *ideal* ensemble function f_{ij}^* should provide the following ranking scores:

$$f_{ij}^* = \begin{cases} 0, & \text{if } c_i = c_j, \\ -1, & \text{if } c_i \neq c_j. \end{cases} \tag{6.22}$$

Using $\boldsymbol{F}^*$ to denote such an ideal ranking score matrix and $\boldsymbol{F}^{ens}$ to denote an estimated score matrix by a given $\boldsymbol{W}$ with Eqn. (6.21), our proposed objective function for metric ensemble learning is then defined as:

$$\rho = \min_{\boldsymbol{W}} \|\boldsymbol{F}^{ens} - \boldsymbol{F}^*\|_F^2 + \nu \mathcal{R}(\boldsymbol{W}), \quad \text{s.t. } \boldsymbol{W} \in S_+^\tau, \tag{6.23}$$

where $\|\cdot\|_F$ denotes a Frobenius norm, and $\mathcal{R}(\boldsymbol{W})$ a regulariser on $\boldsymbol{W}$ with parameter ν controlling the regularisation strength. Whilst common choices of $\mathcal{R}(\boldsymbol{W})$ include L_1, Frobenius norm, or matrix trace, we introduce the following regularisation for a Regularised Metric Ensemble Learning (RMEL) re-id model:

$$\mathcal{R}(\boldsymbol{W}) = -\sum_{i,j} f_{ij}^{ens}, \quad \text{if } c_i = c_j. \tag{6.24}$$

Our intuition is to impose severe penalties for true match pairs with low ranking scores since they deliver the most informative discriminative information for cross-view person re-id, whilst false match pairs are less informative.

Optimisation Eqn. (6.23) is strictly convex with a guaranteed global optimal so it can be optimised by any off-the-shelf toolboxes [160]. We adopt the standard first-order projected gradient descent algorithm [161], with the gradient of Eqn. (6.23) computed as:

$$\nabla_{\boldsymbol{W}} = \sum_{i,j} (f_{ij}^* - f_{ij}^{ens} + \nu \mathcal{I}[c_i = c_j]) \boldsymbol{d}_{ij} \boldsymbol{d}_{ij}^\top, \tag{6.25}$$

with $\mathcal{I}$ being the indicator function. Our optimisation algorithm is summarised in Algorithm 3.

HOL Person Re-Id Given the estimated optimal ensemble weight matrix $\boldsymbol{W}$ and the weak mod-

Algorithm 3: Regularised Metric Ensemble Learning (REML)

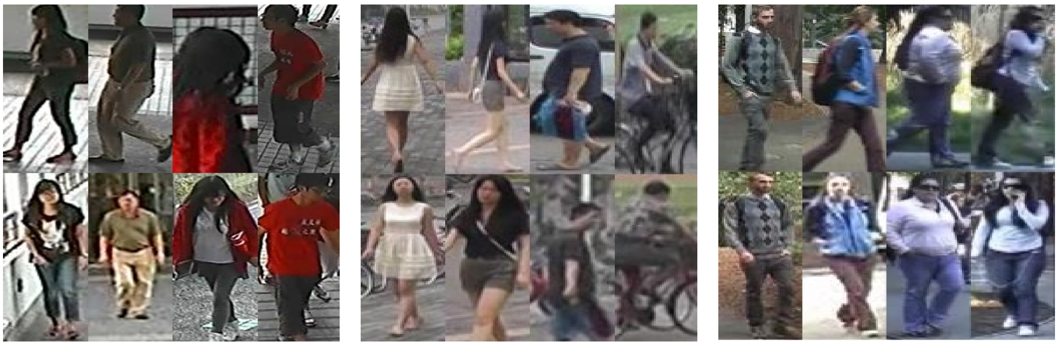
```
Input: Training dataset $\mathcal{X}_{tr} = \{(\boldsymbol{x}_i^p, \boldsymbol{x}_i^g)\}_{i=1}^{N_l}$, label set $\mathcal{C} = \{c_i\}_{i=1}^{N_l}$, learning rate $\varepsilon$, max learning inteartion
        $\tau_{\text{me}}$, and weak HVIL models $\{\boldsymbol{M}_j\}_{j=1}^{\tau}$;
Output: The optimal weight matrix $\boldsymbol{W}$ for the metric ensemble;
Initialisation: Randomly initialise $\boldsymbol{W}_0$ to some PSD matrix.
Metric Ensemble Learning:
for $k = 1 : \tau_{me}$ do
    Calculate gradient $\nabla_{\boldsymbol{W}_{k-1}}$ (Eqn. (6.25));
    Set $\boldsymbol{W}_k = \boldsymbol{W}_{k-1} - \varepsilon \nabla_{\boldsymbol{W}_{k-1}}$;
    Perform eigen-decomposition of $\boldsymbol{W}_k$: $\boldsymbol{W}_k = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$;
    Project $\boldsymbol{W}_k$ back to PSD cone:
    $\boldsymbol{W}_k = \sum_i max(\lambda_i, 0) \boldsymbol{u}_i \boldsymbol{u}_i^\top$.
Return $\boldsymbol{W}$.
```

els $\{\boldsymbol{M}_j\}_{j=1}^{\tau}$, a single strong ensemble model (Eqn. (6.21)) is made available for performing automated HOL re-id of any further probes on the gallery population. Our experiments (Sec. 6.5.2) show that the proposed RMEL algorithm achieves superior performance as compared to state-of-the-art supervised re-id models given the same amount of labelled data.

6.4 Datasets and Experimental Settings

Two sets of comparative experiments were conducted: (1) The proposed HVIL model was evaluated under a *Human-In-the-Loop* (HIL) re-id setting and an *enlarged* test gallery population was used to reflect real-world use-cases (Sec. 6.5.1). (2) In the event of limited human labour budget being exhausted and human feedback becoming unavailable, the proposed HVIL-RMEL model was evaluated under an automated *human-out-of-the-loop* (HOL) re-id setting (Sec. 6.5.2).

Datasets Two largest person re-id benchmarks: CUHK03 [1] and Market-1501 [2], were chosen for evaluations due to the need for large test gallery size. CUHK03 contains 13,164 bounding box images of 1,360 people. Two versions of person image are provided: manually labelled and automatically detected, with the latter presenting more realistic detection misalignment challenges for practical deployments (Fig. 6.5(a)). We used both. Market-1501 has 32,668 person bounding boxes of 1,501 people, obtained by automatic detection. Both datasets cover six outdoor surveillance cameras with severely divergent and unknown viewpoints, illumination conditions, (self)-occlusion and background clutter (Fig. 6.5(b)). In addition, we also selected the most common benchmark VIPeR [7] characterised with low imaging resolution and dramatic illumination variations (Fig. 6.5(c)). Compared to CUHK03 and Market-1501, VIPeR has a much smaller population size (632 people) with fewer (1,264) labelled person images, therefore only suitable for the conventional HOL re-id setting. These three datasets present a wide range of re-id



(a) CUHK03

(b) Market-1501

(c) VIPeR

Figure 6.5: Examples of cross-view person images from three person re-id datasets. Two images in each column describe the same person.

evaluation challenges under different viewing conditions and with different population sizes, as summarised in Table 6.1.

Dataset	Cams	IDs	Labelled	Detected	HIL Split	HOL Split
VIPeR [7]	2	632	1,264	0	-	316/316
CUHK03 [1]	6	1,467	13,164	13,164	1,000	360
Market-1501 [2]	6	1,501	0	32,668	1,000	501

Table 6.1: Settings of three person re-id datasets.

Data Partitions For CUHK03 or Market-1501, we randomly selected 1,000 identities D_{p1} (p stands for population) as the partition to perform *HIL* re-id experiments. The remaining partition of people D_{p2} (360 on CUHK03, and 501 on Market-1501) were separated for evaluating the proposed model against state-of-the-art supervised re-id methods for automated *HOL* re-id (see details in Sec. 6.5.1 and Sec. 6.5.2). Due to its small size, VIPeR was only used in the HOL experiments and the identities were split half-half for training and testing. To obtain statistical reliability, we generated 10 different trials with different random partitions and reported their averaged results.

Visual Features We adopted two types of image features: **(1)** The WHOS descriptor [61]: A state-of-the-art *hand-designed* person re-id feature (5,138 dimensions) composited by colour, HOG [29] and LBP [113] histograms extracted from horizontal rectangular stripes[3]. **(2)** The CNN feature learned by a recently proposed deep architecture for re-id [162]: In contrast to hand-crafted WHOS features, deep CNN features are extracted from a deep model trained by supervised learning from a large number of labelled training data. Specifically, we trained the deep

[3] The LOMO (26,960-D) [24] and GOG (27,622-D) [96] were not selected due to their high dimensionality property which poses high computational cost for online model updating, although they are possibly more discriminative.

model with the entire person search dataset [163], which is independent of CUHK03, Market-1501 and VIPeR, therefore without any additional effect on their data partitions. The trained deep model is directly deployed as a feature extractor (1,024 dimensions) without any domain transfer learning by fine-tuning on the three evaluation datasets. Whilst adopting deep features from training a CNN model using labelled data may seem to be inconsistent with the objective of this work – eliminating the need for offline pre-collected training data, the main purposes of utilising the CNN feature are: (a) To evaluate the proposed HVIL on different features; (b) To demonstrate any additional benefit of the proposed HVIL model on a strong deep feature already learned from a large size labelled training data.

Evaluation Metrics We adopted three performance evaluation metrics in the following experiments: (1) Cumulative Match Characteristic (CMC): calculated as the cumulative recognition rate at each rank position. (2) Expected Rank (ER): defined as the average rank of all true matches. (3) Mean Average Precision (mAP): first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes. For all HIL re-id models, we used the ranking result after the final human feedback applied on each probe. The averaged results over all 10 trials were reported in comparisons.

6.5 Experiments and Evaluations

6.5.1 Human-In-the-Loop Re-Identification Evaluation

Experiment Settings

Probe/Gallery Configuration For each of the D^i_{p1} partitions, we built a probe set for human operators to perform HIL re-id. In each trial, the probe set $\mathcal{P}^i$ contains randomly selected 300 persons with one image/person. For building the cross-view gallery set, we considered three different configurations to fully analyse the behaviour and scalability of the proposed HVIL method:

1. *Single-shot gallery* $\mathcal{G}^i_s$: We randomly selected one cross-view image/person of all the 1,000 identities in partition D^i_{p1} and construct a single-shot gallery set $\mathcal{G}^i_s$ (1,000 person images) on both CUHK03 and Market-1501.

2. *Multi-shot gallery* $\mathcal{G}^i_m$: We built the multi-shot gallery $\mathcal{G}^i_m$ by following [2]. In particular, for all the 1,000 identities in partition D^i_{p1}, we used all cross-view images to construct the

Feature	WHOS [61]									CNN [162] (except for DGD and Inception-V3)								
Dataset	CUHK03 (L)			CUHK03 (D)			Market-1501 (D)			CUHK03 (L)			CUHK03 (D)			Market-1501 (D)		
Rank (%)	1	50	100	1	50	100	1	50	100	1	50	100	1	50	100	1	50	100
L2	2.9	31.1	43.2	2.7	29.8	41.6	16.1	66.6	76.6	19.0	72.0	82.3	17.1	67.0	78.1	44.2	94.4	97.5
kLFDA [23]	5.9	47.3	60.1	4.7	39.6	51.7	21.8	85.8	91.5	21.4	77.4	86.2	19.4	73.7	82.7	52.9	**97.2**	**98.5**
XQDA [24]	3.7	40.2	53.6	2.4	22.4	33.3	18.3	75.1	83.5	19.8	76.9	85.8	17.7	73.9	83.0	49.6	97.0	**98.5**
MLAPG [25]	4.2	39.5	52.4	3.5	36.1	49.3	24.1	84.5	91.2	11.8	69.6	82.5	10.2	64.3	77.9	37.7	95.5	97.9
NFST [47]	7.1	41.5	54.7	4.9	37.4	48.5	34.4	85.3	90.7	9.9	41.7	51.3	9.5	38.0	47.8	45.0	89.7	93.3
HER [95]	7.6	46.0	58.1	5.7	41.8	53.8	39.1	**90.8**	**94.7**	16.2	73.5	84.3	14.5	69.9	80.2	44.0	96.1	98.3
DGD [9]	-	-	-	-	-	-	-	-	-	12.0	58.0	69.8	10.1	49.8	61.6	58.4	95.7	97.4
Inception-V3 [10]	-	-	-	-	-	-	-	-	-	15.7	63.7	74.4	15.3	62.5	72.2	51.6	94.7	96.8
EMR [105]	29.3	29.3	40.7	27.7	27.7	39.5	64.2	64.2	74.2	73.5	73.5	83.7	66.7	66.7	77.5	92.7	92.7	96.8
Rocchio [104]	32.0	38.7	46.2	29.0	36.2	43.8	61.7	70.2	77.5	62.0	79.2	85.2	56.2	74.3	80.8	81.2	94.5	93.3
POP [69]	44.0	51.5	60.0	41.7	48.5	58.8	75.0	78.5	84.5	74.7	74.8	77.2	69.0	70.7	73.2	92.8	93.0	93.3
HVIL (Ours)	**60.2**	**68.2**	**78.5**	**53.7**	**65.0**	**75.3**	**84.5**	89.2	93.2	**84.2**	**89.2**	**93.3**	**80.3**	**86.0**	**91.2**	**95.3**	96.0	98.3

Table 6.2: Human-in-the-loop person re-id with **single-shot** galleries. Gallery Size: 1,000 for both CUHK03 and Market-1501; L: Labelled; D: Detected.

gallery set. As such, the average gallery size is 4,919 on CUHK03 and 9,065 on Market-1501. Note that, we did not utilise the label information about which images are of the same person, and thus both CMC and mAP can be used for performance evaluation.

3. *Open-world gallery* $\mathcal{G}_d^i$: We considered a more challenging setting with a large number of distractors involved in the gallery set. Specifically, we added 34,574 bounding boxes of 11,934 persons from the person search dataset [163] to the single-shot gallery set $\mathcal{G}_s^i$. The resulted gallery $\mathcal{G}_d^i$ size is 35,574 on both datasets. This is to evaluate the scalability of HIL re-id methods when operating under the open-world re-id setting featured with a huge gallery search space.

Human Feedback Protocol Human feedback were collected on all 10 trials of D_{p1}^i partitions and all 3 different gallery configurations, in total $3 \times 10 = 30$ independent sessions on each dataset by 5 volunteers as users. During each session, a user was asked to perform the *HIL* re-id on probes in probe set $\mathcal{P}^i$ against gallery set $\mathcal{G}^i \in \{\mathcal{G}_s^i, \mathcal{G}_m^i, \mathcal{G}_d^i\}$. For each probe person, a *maximum* of 3 rounds of user interaction are allowed. We limited the users to verify only the top-50 in the rank list (5% of $\mathcal{G}_s^i$, $0.5 \sim 1\%$ of $\mathcal{G}_m^i$, and 0.1% of $\mathcal{G}_d^i$). During each interaction: (1) A user selects one gallery image as either *strong-negative* or *true-match*; and (2) the system takes the feedback, updates the ranking function and returns the re-ordered ranking list, all in real-time (Sec. 6.2). The HVIL model was evaluated against eight existing models for *HIL* re-id deployment as follows.

HIL Competitors Three existing HIL models were compared: (1) POP [69]: The current state-of-the-art HIL re-id method based on Laplacian SVMs and graph label propagation; (2) Roc-

Feature	WHOS [61]						CNN [162] (except for DGD and Inception-V3)					
Dataset	CUHK03 (L)		CUHK03 (D)		Market-1501 (D)		CUHK03 (L)		CUHK03 (D)		Market-1501 (D)	
Rank (%)	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
L2	4.1	14.1	3.6	13.9	28.0	23.9	22.0	29.5	20.7	28.0	58.0	50.9
kLFDA [23]	8.1	17.8	6.3	16.5	47.1	39.9	25.4	32.8	23.9	31.0	67.7	63.0
XQDA [24]	3.6	14.9	4.5	14.5	34.3	30.1	24.5	31.7	22.5	30.0	63.4	58.1
MLAPG [25]	5.0	15.1	5.1	15.1	44.3	40.8	14.8	23.7	12.2	21.9	54.5	50.8
NFST [47]	8.2	17.5	7.7	16.6	68.3	62.1	20.2	26.8	18.6	25.3	76.2	69.9
HER [95]	9.5	18.6	8.1	17.4	68.9	61.7	24.3	31.4	22.3	29.3	77.4	72.1
DGD [9]	-	-	-	-	-	-	15.1	23.5	13.0	21.4	82.1	75.9
Inception-V3 [10]	-	-	-	-	-	-	19.2	27.1	18.3	26.2	76.3	71.4
EMR [105]	30.8	20.2	29.7	19.3	76.0	31.7	71.3	40.6	66.3	37.5	94.0	57.7
Rocchio [104]	34.0	26.4	30.7	23.7	74.3	37.1	59.3	50.0	56.0	46.8	83.7	65.1
POP [69]	43.0	39.4	44.3	38.2	82.7	52.7	71.7	68.2	68.0	64.3	94.0	74.0
HVIL (Ours)	**63.0**	**59.0**	**53.7**	**48.7**	**87.3**	**63.3**	**84.0**	**73.4**	**80.7**	**72.7**	**96.0**	**83.3**

Table 6.3: Human-in-the-loop person re-id with **multi-shot** galleries. Gallery Size: 4,919 for CUHK03 and 9,065 for Market-1501. L: Labelled; D: Detected.

chio [104]: A probe vector modification model updates iteratively the probe's feature vector based on human feedback, widely used for image retrieval tasks [164]; (3) EMR [105]: A graph-based ranking model that optimises the ranking function by least square regression. For a fair comparison of all four HIL models, the users were asked to verify the same probe and gallery data $(\mathcal{P}^i, \mathcal{G}^i)$ with the same two types of feedback given the ranking-list generated by each model.

HOL Competitors In addition, seven state-of-the-art conventional HOL supervised learning models were also compared: kLFDA [23], XQDA [24], MLAPG [25], NFST [47], HER [95], DGD [9], and Inception-V3 [10], among them two are deep learning models (DGD and Inception-V3). These supervised re-id methods were trained using fully pre-labelled data in the separate partition D^i_{p2} (CUHK03: averagely 3,483 images of 360 identities; Market-1501: averagely 7,737 images of 501 identities) before being deployed to $\mathcal{P}^i$ and $\mathcal{G}^i$ for automated HOL re-id testing. Note, the underlying human labour effort for pre-labelling the training data to learn these supervised models was significantly greater – exhaustively searching 3,483 and 7,737 *true* matched images respectively for CUHK03 and Market-1501, than that required by the HIL methods – between 300 to 900 *indicative* verification (strong negative or true match) given a maximum of 300 probes on both CUHK03 and Market-1501, so only 1/10th of and weaker user input than supervised HOL models. It should be noted that non-deep distance metric models (kLFDA, XQDA, MLAPG, NFST, HER) were trained using either hand-crafted WHOS [61] or deep learning CNN [162] features (Section 6.5), while DGD and Inception-V3 were trained directly from raw images in D^i_{p2}, since these two deep models provide their own deep CNN features (256 dimensions for DGD and 2,048 for Inception-V3).

Implementation Details For implementing the HVIL model (Sec. 6.2), the only hyper-parameter

Feature	WHOS [61]									CNN [162] (except for DGD and Inception-V3)								
Dataset	CUHK03 (L)			CUHK03 (D)			Market-1501 (D)			CUHK03 (L)			CUHK03 (D)			Market-1501 (D)		
Rank (%)	1	50	100	1	50	100	1	50	100	1	50	100	1	50	100	1	50	100
L2	2.8	27.2	38.2	2.6	24.8	34.4	10.7	43.9	51.5	18.3	69.6	80.1	16.6	65.0	75.9	31.4	77.6	84.0
kLFDA [23]	5.6	32.9	44.8	3.6	28.1	38.0	19.8	67.6	76.1	17.7	66.9	77.1	16.8	63.6	72.9	38.4	84.2	89.7
XQDA [24]	3.1	25.3	36.3	2.4	21.7	32.0	16.6	61.9	70.7	15.5	61.7	70.6	13.2	58.1	67.7	31.3	77.0	84.4
MLAPG [25]	3.7	33.3	44.0	2.8	28.9	39.2	18.9	67.6	76.3	6.4	34.6	43.1	5.8	30.2	37.9	20.1	65.0	74.0
NFST [47]	5.6	34.6	45.6	4.2	30.3	40.5	30.1	78.3	85.0	9.8	41.4	51.0	9.4	37.8	47.5	39.6	83.7	88.4
HER [95]	6.3	36.2	46.0	4.5	31.4	40.5	32.7	80.8	86.0	12.3	57.3	66.5	11.8	54.7	64.1	26.1	70.6	79.0
DGD [9]	-	-	-	-	-	-	-	-	-	7.2	29.1	35.0	5.6	23.2	29.0	48.6	86.3	89.2
Inception-V3 [10]	-	-	-	-	-	-	-	-	-	8.9	31.5	38.2	7.4	30.5	37.6	37.0	79.4	83.9
EMR [105]	25.8	25.8	35.5	23.1	23.1	32.2	40.8	40.8	46.8	70.7	70.7	81.0	66.3	66.3	77.7	72.7	72.7	80.7
Rocchio [104]	28.7	32.3	37.5	25.3	30.0	37.0	43.6	46.2	48.8	61.0	74.0	81.7	56.7	73.7	80.0	64.3	74.3	80.0
POP [69]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HVIL (Ours)	**55.6**	**65.7**	**74.8**	**52.0**	**60.3**	**67.8**	**61.7**	**70.8**	**76.7**	**80.3**	**86.0**	**91.3**	**73.3**	**84.7**	**89.3**	**91.3**	**93.3**	**96.0**

Table 6.4: Human-in-the-loop person re-id with **open-world** galleries consisting of 34,574 **distractors**. Gallery Size: 35,574 for both CUHK03 and Market-1501. L: Labelled; D: Detected. Note: POP results are unavailable because it was *intractable* on our computing hardware.

η (Eqn. (6.5)) was set to 0.5 on both CUHK03 and Market-1501. We found that HVIL is insensitive to η with a wide satisfiable range from 10^{-1} to 10^{1}. For POP, EMR, and Rocchio, we adopted the authors' recommended parameter settings as in [69, 104]. For all HIL methods above, we applied L_2 distance as the initial ranking function $f_0(\cdot)$ without loss of generalisation[4]. Note that for HVIL, once $f_0(\cdot)$ was initialised for only the very first probe, it was then optimised incrementally across different probes. In contrast, for POP and EMR and Rocchio, each probe had its own $f_0(\cdot)$ initialised as L_2 since the models are not cumulative across different probes. For HOL competitors, the parameters were determined by cross-validation on D_{p2} with the authors' published codes. All models except DGD and Inception-V3 used the same two feature descriptors for comparison (WHOS [61] and CNN feature [162]). DGD [9] and Inception-V3 [10] used their own deep features from training their CNN networks.

Evaluations on Person Re-Identification Performance

The person re-id performances of all HIL and HOL methods on $\mathcal{P}^i$ and $\{\mathcal{G}_s^i, \mathcal{G}_m^i, \mathcal{G}_d^i\}$ are compared in Tables 6.2 (single-shot), 6.3 (multi-shot), and 6.4 (open-world) respectively.

HIL vs. HOL Re-Id Methods We first compared the re-id matching performance of HIL and HOL re-id schemes. It is evident from the three Tables that the HIL methods outperform significantly the conventional HOL counterparts in all testing settings on both datasets. Specifically, in single-shot setting (Table 6.2), *all* conventional supervised re-id models suffered severely when the gallery size was enlarged to 1,000 from their standard setting. For example, the state-of-the-

[4] No limitation on considering any other distance or similarity metrics, either learned or not. However, non-learning based generic metrics are more scalable and transferable in real-world.

art deep re-id model DGD [9] can achieve 72.6% Rank-1 rate on CUHK03 (Labelled) under the test protocol of using the 100-sized test gallery. However, its Rank-1 accuracy drops dramatically to only 12.0% Rank-1 on CUHK03 (Labelled) and 10.1% (Detected) under the 1,000-sized test gallery evaluated here. Similar performance drops occur for all other HOL models. Such low Rank-1 matching accuracies show that, existing best supervised re-id approaches are still far from being sufficiently mature to provide a fully automated HOL re-id solution in real world. On the contrary, HIL methods make more realistic assumptions by considering human in the loop, and leverage limited human efforts to directly drive up model matching performance by mining the joint human-machine benefits. The advantage in re-id matching by the HOL methods is clear: for example, with WHOS feature the proposed HVIL achieves over 50% and 80% in Rank-1 on CUHK03 and Market-1501 (Table 6.2), which is much more acceptable in practical use. In terms of supervision cost, the supervised HOL models were offline trained on a large-sized pre-labelled data in D_{p2} with an average of 3,483 cross-view images of 360 identities on CUHK03, and 7,737 images of 501 identities on Market-1501. Whereas the HIL models required much less human verification effort, e.g. at most 3 feedback for each probe in top-50 ranks only, in total ($300 \sim 900$) weak feedback. Human feedback is neither restricted to be only true matches, nor exhaustively labelling person identity labels, nor searching true matches in a huge image pool. These evidences suggest that HIL re-id is a more cost-effective and promising scheme in exploiting human effort for real-world applications as compared to the conventional HOL approach.

Among all HIL re-id models, the proposed HVIL achieves the best performance. For instance, it is found in Table 6.2 that the HVIL improves significantly over the state-of-the-art HIL model POP on Rank-1 score, e.g. from 44.0% to 60.2% on CUHK03 (Labelled), from 41.7% to 53.7% on CUHK03 (Detected), and from 75.0% to 84.5% on Market-1501, when the WHOS feature is used. HVIL's advantage continues over all ranks. This demonstrates the compelling advantages of the HVIL model in cumulatively exploiting human verification feedback, whilst other existing human-in-the-loop models have no mechanisms for sharing human feedback knowledge among different probes.

Effect of Features Next, we evaluated the effect of different visual features by comparing the hand-crafted WHOS [61] and the most recent deep CNN feature [162] learned from the large scale person search dataset [163]. As shown in Table 6.2, the CNN feature is much more discriminative and view-invariant than the WHOS thanks to the access of large quantity of labelled data and the strong deep model learning capacity. Specifically, with CNN feature, even the generic L2 metric can achieve 19.0%/17.1% and 44.2% on CUHK03 (Labelled/Detected) and Market-1501, respectively. Importantly, CNN feature can be well complementary with HIL re-id methods: The HIL re-id Rank-1 rates are further boosted to a more satisfying level, e.g. 84.2%/80.3% and 95.3% by the proposed HVIL. This implies the great compatibility of the HVIL with deep feature learning. On the other hand, it is found that with such a powerful deep CNN feature, HOL models are still outperformed drastically by HIL methods. This suggests the consistent and general advantages of the HIL re-id scheme over the HOL approach given various types of visual features.

Single-Shot vs. Multi-Shot We evaluated the effect of shot number in the gallery set in person re-id performance. When more shots of a person are available (Table 6.3 vs. Table 6.2), re-id matching accuracy can be improved in most cases by either HIL and HOL methods including the proposed HVIL. However, the best results are still generated by the HVIL model. This suggests the steady advantage of the proposed method in different search gallery settings. In particular, we have the following observations and justifications: (1) The Rank-1 improvement degree varies over different datasets, with Market-1501 benefiting more than CUHK03. The plausible reason is that, Market-1501 person images give more pose and detection misalignment challenge due to poorer person bounding box detection, and therefore multi-shot images with various poses and detection qualities can bring more gains. (2) The HVIL model seem to benefit less from multi-shot gallery images as compared to other methods. This may be due to the better capability of mitigating the pose/detection misalignment challenge by the proposed incremental model learning, thus not needing multiple shots as much as the other models do.

Effect of Distractors in Open-World Setting Finally, we evaluated the effect of open-world distractors in the gallery set for further testing the model scalability. This evaluation is made by comparing Table 6.2 and Table 6.4. After adding 34,574 person bounding boxes as distractors to the 1000 sized single-shot gallery (i.e. the gallery size is enlarged by 35 times), we observed that (1) As expected, all methods suffered from some drop in re-id performance; (2) The HIL methods outperform more significantly the HOL models under the open-world setting; and (3) the HVIL again achieves the best re-id performance, and particularly on the CUHK03 (Detected) dataset, the addition of 34K distractors causes only a $1.7\% = 53.7 - 52.0\%$ Rank-1 drop. This again

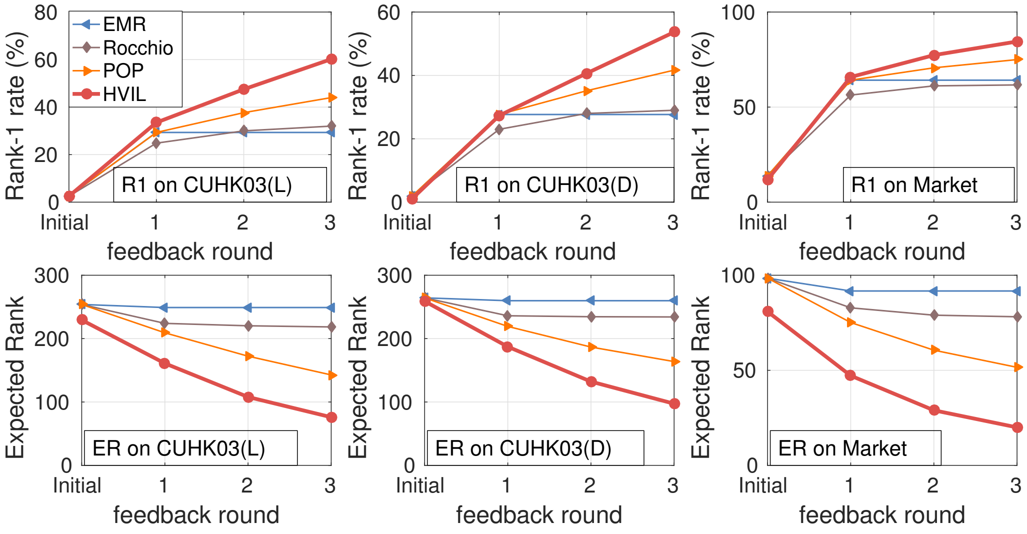


Figure 6.6: Comparing Rank-1 score and Expected Rank (ER) on human feedback rounds.

suggests the clear advantages and superiority of having human in the loop for real-world person re-id applications when the gallery population size is inevitably large in the open-world operation scenarios. More specifically, when the WHOS feature was used, the best HOL model HER's Rank-1 rates dropped from 7.6% to 6.3%, 5.7% to 4.5%, and 39.1% to 32.7% on CUHK03 (Labelled), CUHK03 (Detected), and Market-1501 respectively. The best HIL competitor, POP, completely fails to operate with such a large gallery set. The reason is that POP requires to build an affinity graph and calculate the graph Laplacian on all the gallery samples to propagate human labels. Given a 34,574-sized gallery set, the affinity graph alone takes 4.78 GB storage which is both difficult to process (out of memory) for common workstations and suffering from slow label propagation.

Further Analysis on Human Verification

We examined the effectiveness of the proposed HVIL model in exploiting human verification effort for HIL re-id in the single-shot setting with WHOS feature.

Statistics Analysis on Human Verification Fig. 6.6 shows the comparisons of Rank-1 and Expected Rank (ER) on the 4 human-in-the-loop models over three verification feedback rounds. It is evident that the proposed HVIL model is more effective than the other three models in boosting Rank-1 scores and pushing up true matches' ranking orders. The reasons are: (1) Given a large gallery population with potentially complex manifold structure, it is difficult to perform accurately graph label propagation for graph-based methods like POP and EMR. (2) Unlike POP/EMR/Rocchio, the proposed HVIL model optimises on re-id ranking losses (Eqn. (6.3))

Dataset	CUHK03 (L)			CUHK03 (D)			Market-1501 (D)		
Method	HVIL	POP	ES	HVIL	POP	ES	HVIL	POP	ES
Found-matches(%) ↑	60.2	44.0	**100**	53.7	41.7	**100**	84.5	75.0	**100**
Browsed-images ↓	**35.1**	57.3	253.9	**71.6**	107.0	264.3	**19.7**	33.8	98.5
Feedback ↓	**2.2**	2.4	-	**2.4**	**2.4**	-	**1.6**	1.7	-
Search-time(sec.) ↓	**23.5**	47.3	187.0	**33.0**	55.8	234.9	**14.7**	22.7	131.8

Table 6.5: Human verification effort vs. benefit. All measures are from averaging over all probes. ↓: lower better; ↑: higher better. Setting: single-shot. Feature: WHOS.

specifically designed to maximise the two types of human verification feedback. (3) The HVIL model enables knowledge cumulation (Eqn. (6.5)). This is evident in Fig. 6.6 where HVIL yields notably better (lower) Expected Ranks (ER), even for the initial ER before verification feedback takes place on a probe (due to benefiting cumulative effect from other probes). In contrast, other models do not improve initial ER on each probe due to the lack of a mechanism to cumulate experience.

Human Verification Cost-Effectiveness We further evaluated the human verification effort in relation to re-id performance benefit by analysing the meta statistics of HIL re-id experiments above. We compared the HVIL model with the POP model and Exhaustive Search (ES) where a user performs exhaustive visual searching over the whole gallery ranking list (1,000) generated by L2 metric until finding a true match. The averaged statistics over all 10 trials were compared in Table 6.5. It is evident that though ES is guaranteed to locate a true match for every probe if it existed, it is much more expensive than POP (3×) and HVIL (5×) in search time given a 1,000-sized gallery. This difference will increase further on larger galleries. Comparing HVIL and POP, it is evident that HVIL is both more cost-effective (less Search-time, Browsed-images and Feedback) and more accurate (more Found-matches).

HIL Re-Id Search Speed To better understand model convergence given human feedback, we conducted a separate experiment to measure the search time by different human-in-the-loop models given the initial rank lists on 25 randomly selected probes verified by multiple users. This experiment was evaluated by 10 independent sessions with the same set of 25 probes provided. In each session, the users were required to find a true match for all 25 probes. Specifically, for HVIL and POP, if a true match was not identified after 3 (maximum) feedback, the users then performed an exhaustive searching until it was found. The search time statistics for all 25 probes are shown in Fig. 6.7, where a bar shows the variance between 10 different sessions. It is

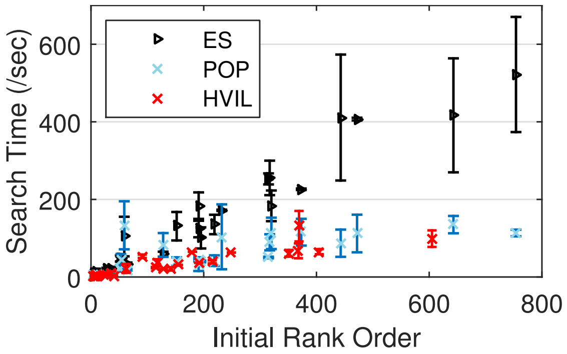


Figure 6.7: Search time from different HIL models on the same 25 randomly selected probes. Dataset: CUHK03 (Labelled). Setting: single-shot.

unsurprising that ES is the least efficient whilst HVIL is the quickest in finding a true match, i.e. the data points of HVIL are much lower in search time. Moreover, it is evident that HVIL yields much better initial ranks, i.e. the data points of HVIL are more centred towards the bottom-left corner. This further shows the benefit of cumulative learning in HVIL (Sec. 6.2.3).

Strong vs. Weak Negative We evaluated explicitly the effect of strong and weak negative feedback on the HIL re-id performance. To this end, a further experiment was conducted with the single-shot gallery setting with WHOS feature, under the same human feedback protocol as described in Sec. 6.5.1 with the only difference that users were required to label visually similar samples (weak negative) instead of dissimilar ones (strong negative). For model updates on weak negatives we adopted the same loss design of our preliminary model [48]. Table 6.6 shows that labelling weak negatives is much less effective than strong ones in re-id performance. For example, when weak negatives are labelled instead of strong ones, Rank-1 rates drops from 60.2%/53.7%/84.5% to 45.3%/43.6%/78.0% and Expected Ranks increases from 76.0/99.8/20.0 to 203.0/226.7/90.7 on CUHK03(Labelled/Detected) and Market-1501. Moreover, it is indicated by the users that weak negatives are much harder and time consuming to label. This is intuitive given that most top-ranked gallery images are visually similar which renders a user hard to select a specific one against the others (Fig. 6.3c).

Dataset	CUHK03 (L)		CUHK03 (D)		Market-1501 (D)	
Metric	R1(%)	ER	R1(%)	ER	R1(%)	ER
Strong	**60.2**	**76.0**	**53.7**	**99.8**	**84.5**	**20.0**
Weak	45.3	203.0	43.6	226.7	78.0	90.7

Table 6.6: Effect of strong and weak negatives in HIL re-id performance.

6.5.2 Human-Out-of-the-Loop Re-Identification Evaluation

Experiment Settings

Finally, we assume that a limited budget for human verification on D^i_{p1} has been reached after time τ so that human feedback becomes unavailable. Re-id of any further independent population (e.g. D^i_{p2}) turns to a conventional human-out-of-the-loop (HOL) re-id problem, if one treats previously human labelled samples as training samples. The proposed RMEL model was then evaluated under this HOL re-id setting against both state-of-the-art supervised models and baseline ensemble models. This experiment was conducted with CNN feature on both CUHK03 (Labelled) and Market-1501 dataset, Additionally, to examine our proposed HVIL-RMEL framework in a more comparable context defined in the literature on HOL re-id, we also tested on the VIPeR [7] benchmark, with more details as follows.

Training/Testing protocol On CUHK03 and Market-1501 datasets, in each of the overall 10 trials, we employed the human verified true matches on D^i_{p1} to learn the weights for constructing a strong ensemble model using all the verified weak models $\{\boldsymbol{M}_j\}^{\tau}_{j=1}$ collected from our previous experiments on human-in-the-loop re-id. The strong ensemble model was then deployed for testing on the separate partition D^i_{p2} with the size of 360 and 501 persons for CUHK03 and Market-1501 respectively. For performance evaluation, we adopted the standard single-shot test setting, i.e. randomly sampling 360 cross-camera person image pairs from CUHK03 and 501 pairs from Market-1501 on $\{D^i_{p2}\}^{10}_{i=1}$ to construct the test gallery and probe sets over ten trials. On VIPeR dataset, we followed the exact setting of the established protocol in existing literature: splitting the 632 identities into $50-50\%$ partitions for training and testing sets. For obtaining weak re-id models, we simulated HVIL feedback update by simply giving the ground-true matching pairs instead of strong negatives (Eqn. (6.12)); therefore each weak model was obtained by a true-match, using the same information as training a conventional supervised model. On all three datasets the averaged CMC performance over all trials was reported.

HOL Competitors On CUHK03 and Market-1501, five state-of-the-art supervised re-id models are compared: kLFDA [23], XQDA [24], MLAPG [25], NFST [47], HER [95] were trained using 300 ground-truth labelled data from $\mathcal{P}^i$ (300) and $\mathcal{G}^i_s$ (1,000) of D^i_{p1} under the same CNN feature, for both CUHK03 and Market-1501. The trained models were tested on the separate

[5] In this study, a challenging single-shot training/testing protocol (300/360 for CUHK03 and 300/501 for Market-1501) is adopted for HOL evaluation (Table 6.7). In contrast to the reported multi-shot setting [1, 2] of 1260/100 for CUHK03 and 751/750 for Market-1501, this is a harder task.

Dataset	CUHK03 ($N_g = 360$)				Market-1501 ($N_g = 501$)			
Rank (%)	1	5	10	20	1	5	10	20
kLFDA [23]	20.6	43.1	55.8	67.8	57.0	83.9	91.9	96.9
XQDA [24]	19.7	43.6	56.7	68.9	52.9	83.5	89.9	96.1
MLAPG [25]	15.8	35.8	45.6	57.7	52.2	78.6	87.7	94.1
NFST [47]	22.8	43.1	56.1	63.7	58.6	84.1	90.7	96.3
HER [95]	**25.3**	43.3	55.8	67.1	60.6	83.9	90.7	96.8
HVIL - $\boldsymbol{M}_{avg}$	19.7	39.2	55.3	70.3	57.3	85.5	93.0	**96.5**
HVIL - $\boldsymbol{M}_{\tau}$	20.3	43.3	56.4	66.1	59.3	86.8	**93.6**	**96.5**
HVIL - RMEL	21.9	**46.7**	**59.2**	**71.4**	**62.6**	**87.0**	92.3	96.3

Table 6.7: Automatic person re-id (HOL) with CMC performances on CUHK03 and Market-1501. Gallery Size: 360 for CUHK03 and 501 for Market-1501^5.

partition D^i_{p2} with same testing protocol as above. On VIPeR, as our training/testing protocol is standard, we compared fifteen recently published state-of-the-art including six deep models: RDC-Net[56], JRL [133], DGD [9], Gated S-CNN [52], S-LSTM [63], MCP [57], and nine shallow models: MLF [21], kLFDA [23], SCNCD [45], XQDA [24], MLAPG [25], RKSL [132], NFST [47], LSSCDL [49], HER [95]. Since most of the above work were reported with the same training/testing protocol but various features and unavailable code access, we simply compared ours with their published results.

Metric Ensemble Baselines For investigate the effect of RMEL's learned ensemble, two baseline models are compared: (1) HVIL - $\boldsymbol{M}_{\tau}$: The incrementally optimised re-id model $\boldsymbol{M}_{\tau}$ obtained by HVIL from the last probe image at time τ during the *human-in-the-loop* process. (2) HVIL - $\boldsymbol{M}_{avg}$: An naive approach to ensemble weak models, that is, simply taking an average weighting of all weak models $\{\boldsymbol{M}_j\}_{j=1}^{\tau}$ as the ensemble re-id model.

Evaluations on Person Re-Identification Performance

Tables 6.7 and 6.8 report the result. For CUHK03, there is insufficient labelled data for all camera pairs during training, given only one pair of randomly selected single-shot images per identity. All models generated poor re-id performances (Rank-1 rates $< 30\%$), much less than state-of-the-art reported in the literature. For Market-1501, a similar problem exists although less pronounced. Note, the results in Table 6.7 are based on a single-shot test setting. This is a much harder problem than the multi-shot test setting [2] where on average 14.8 true matches exist in the gallery for each probe. Given the experimental results above, it is evident that: Due to (1) a much larger unlabelled test gallery population than the labelled training set, (2) a lack of

Dataset	VIPeR ($N_g = 316$)			
Rank (%)	1	5	10	20
MLF [21]	29.1	52.3	66.0	79.9
kLFDA [23]	38.6	69.2	80.4	89.2
SCNCD [45]	33.7	62.7	74.8	85.0
XQDA [24]	40.0	68.1	80.5	91.1
MLAPG [25]	40.7	69.9	82.3	92.4
RKSL [132]	40.2	74.5	**85.7**	**93.5**
NFST [47]	42.3	71.5	82.9	92.1
LSSCDL [49]	42.7	-	84.3	91.9
HER [95]	45.1	74.6	85.1	93.3
RDC-Net[56]	40.5	60.8	70.4	84.4
JRL [133]	38.4	69.2	81.3	90.4
DGD [9]	38.6	-	-	-
Gated S-CNN [52]	37.8	66.9	77.4	-
S-LSTM [63]	42.4	68.7	79.4	-
MCP [57]	**47.8**	**74.7**	84.8	91.1
HVIL - $\boldsymbol{M}_{avg}$	40.8	66.1	76.9	86.4
HVIL - $\boldsymbol{M}_{\tau}$	42.1	69.0	78.5	88.6
HVIL - RMEL	47.1	71.7	82.5	91.3

Table 6.8: Automatic person re-id (HOL) with CMC performances on VIPeR.

sufficient multi-shot training/testing data in many camera pairs, *human-in-the-loop* approach to re-id is not only desirable, but essential for re-id in real world applications.

Nevertheless, for HOL re-id, the proposed HVIL-RMEL still achieves the best performance among all models with a Rank-1 of 21.9% on CUHK03 and 62.6% on Market-1501. More importantly, even though less true-match data (253 pairs for CUHK03 and 285 pairs for Market-1501) were used to learn the ensemble weighting for the RMEL model as compared to the ground-truth data (300 pairs for both benchmarks) used to train kLFDA, XQDA and MLAPG, it is evident that the human verification feedback process yields more discriminative information for optimising probe re-id directly in the gallery population, resulting in a more optimal ensemble model. When HVIL-RMEL was evaluated under the standard training/testing setting on VIPeR, it yields 47.1% for Rank-1 rate, which is only 0.6% lower compared to the current best deep model MCP [57]. It is also evident that naively taking an average ensemble model (HVIL - $\boldsymbol{M}_{avg}$) gives even poorer performance than the cumulatively learned single model (HVIL - $\boldsymbol{M}_{\tau}$).

6.6 Summary

We formulated a novel approach to human-in-the-loop person re-id deployment by introducing a Human Verification Incremental Learning (HVIL) model, designed to overcome two unrealistic assumptions adopted by existing re-id models that prevent them to be scalable to real world applications. In particular, the proposed HVIL model avoids the need for collecting off-line pre-labelled training data and is scalable to re-id tasks in large gallery sizes. The advantage of HVIL over other human-in-the-loop models is its ability to learn cumulatively from human feedback on more probe images when available. We further developed a regularised metric ensemble learning (RMEL) method to explore HVIL for automated re-id tasks when human feedback is unavailable. Extensive comparisons on the CUHK03 [1] and the Market-1501 [2] benchmarks show the potentials of the proposed HVIL-RMEL model for real-world re-id deployments.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis has explored a wide range of approaches to reduce the human labelling efforts in modelling person re-identification, and meanwhile maximise its cost-effectiveness for more scalable model training and deployments (Figure 1 in Abstract). In particular, the primary aims of this thesis are (i) to extract discriminative information from unlabelled surveillance images for model training since they are much easier and cheaper to collect and larger in scales; (ii) to efficiently utilise limited human labelling labour for model training, so that annotation efforts are only concentrated on a small group of data which contributes most to a re-identification model's discriminative power; (iii) to facilitate human operators and speed up the searching time for model deployments with a potentially large searching space. Specifically,

1. We have adopted unsupervised learning based approaches to (i) in Chapter 3 and Chapter 4. Particularly, Chapter 3 proposes a subspace learning model to exploit the inter/intra-view affinity information from *unlabelled* data and learns an efficient closed-form global similarity matching function. On the other hand, Chapter 4 proposes a generative topic model to learn the localised appearance saliency from each unlabelled individual person images, and perform re-identification based on the salient visual features which are representative for each person and robust across camera views. The capability of learning from *unlabelled data* substantially reduces the demand of heavy human labelling for model training, and improves the *scalability* of a re-id model.

2. We have adopted active learning algorithms with incremental model updates on-the-fly to address (ii) in Chapter 5. A new active learning algorithm is proposed for cost-effective human labelling and model update, by only querying the most informative rather than randomly sampled data from a human operator. The active learning model jointly explore the population diversity and discover the discriminative class boundary of the up-to-date model, so that best re-identification matching with the least labelling cost.

3. To address (iii), Chapter 6 has proposed a a new human-in-the-loop re-id model which incrementally adapts its model parameters continuously improves the re-identification retrieval results for each query image by taking only a handful of weak human feedback, without the need for exhaustive eyeball search of true/false in the entire very large gallery set. The model can be directly deployed without the need of heavy human labelling for the pre-collection a separate training dataset.

Although presented as separate chapters in this thesis, techniques proposed in Chapter 3, 4, 5 and 6 should be treated as synergistic building blocks required by one practical re-identification system with remarkably reduced human annotation efforts. In such a system, an operator only needs to annotate a small portion of person images which are automatically selected by a re-identification model. The model could immediately adapt its parameters with any incoming labels, and meanwhile learns complementary information from the vast amount of unlabelled data to further increase its discriminative performance; Given a query image of any individual of interest, an operator does not needs to exhaustively browse and verify every instance on the retrieved image list page by page, instead he/she can take two or three mouse clicks so that a true match will quickly show up in sight. In sum, this thesis considered many aspects to reduce the human efforts involved at different stages of a real-world surveillance system.

7.2 Future Work

The potential research directions for future work beyond the proposed methods are summarised as follows:

- Chapter 3 attempts to increase the scalability of a re-identification model by unsupervised learning from unlabelled data and relaxing the requirement of human labels. To do so it formulated a canonical correlation analysis [112] (CCA) based model, which learns a separate projection/representation for every different camera view i.e. view-specific representation models, which could be harmful to the approach's scalability. In a real-scale surveillance network where the camera numbers increases dramatically and candidate images could come from multiple or even unknown views, it is more desirable to train a unified model which is independent to camera views. One possible direction could be formulate other base frameworks, such as Linear Discriminant Analysis [121], metric learning [138], or deep learning [10], into unsupervised learning models by imposing the intra/inter-view affinity constraints (Section 3.3) as learning principles.

- In Chapter 4, although the proposed generative topic model successfully learns the salient appearance regions on unlabeled images, the extracted saliency map and foreground/background maps are still coarse (see Figure 4.3). The reason is that Chapter 3 adopts a patch-based representation, where a single saliency score will be assigned to all pixels within the same image patch. In recent years many end-to-end pixel level saliency mapping techniques has been proposed, e.g. [165, 166, 167]. However, unlike Chapter 4, these above mentioned approaches are supervised models and require even heavier and more fine-grained human annotations (pixel level). One potential future work is to exploit the idea proposed in Chapter 4 with the more advanced pixel-level models for accurate saliency generation under the unsupervised learning setting, and then achieve more discriminative re-identification matching results.

- Chapter 5 leaves a few open questions to exploiting active learning and incremental learning algorithms in the context of re-identification. First, what forms of data input should an active learning model ask for human labeling? In the current approach proposed by Chapter 5, an active learning model automatically selects one unlabeled image, and ask for human operators to label its cross-view matching pair. In real-world scenarios, this is however not guaranteed (see Section 1.2). When a true match cannot be labeled, the model takes no inputs and human annotation effort will be wasted. One possible solution is to develop active learning algorithms which are capable of taking more flexible types of human inputs, e.g. the weak human feedback as proposed in Chapter 6 and [69], for a model to perform update. A second question is that, is it possible to achieve fast incremental updates for more advanced models such as deep neural networks [10]? Currently the incremental model proposed by Chapter 5 is characterised by its closed-form solution and efficient updates, but as a shallow regression model it sacrifice discriminative capability compared to recent deep models [9, 50, 52]. However, it is widely known that deep neural networks require iterative stochastic gradient decent optimisations on a batch of data to adapt its parameters, which is both inefficient in timing and less effective for updates with single data. One interesting and yet unsolved problem is thus to perform fast incremental model updates with deep learning based models on a steam of incoming data.

- Currently, Chapter 6 treats human operators as an adversary in the proposed online learning system, which generates the loss at each time frame based on his/her feedback. Although shown to be effective, e.g. with dramatically reduced annotation effort and boosted re-identification accuracy, the approach can still be further improved in many aspects. For example, the current system overwhelmingly relies on the correctness of a human operator, whereas in real-world a human could easily make mistakes which cause the model parameters to converge to an unexpected state. Moreover, the quality of feedback depends on many uncontrolled factors such as experience, concentration, mental and physical condition during working, etc. One possible solution to avoid this is to explore recently developed reinforcement learning (RL) algorithms [168, 169] where a RL agent could automatically learn from its past experiences without the need for explicit human input. In addition, more sophisticated adversary mechanism and reward functions should be also designed to further reduce the involvement of human, in order to achieve automated retrieval results refinement.

Appendix A

Derivation of FDA Coding

In the following, we provide a detailed derivation of FDA coding (Eq. (5.4)) in our IRS method.

FDA Criterion. Specifically, the FDA criterion aims to minimise the intra-class (person) appearance variance and maximise inter-class appearance variance. Formally, given zero-centred training data $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$, we generate three scatter matrices defined as follows:

$$\begin{aligned}
\boldsymbol{S}_w &= \frac{1}{n}\sum_{j=1}^{c}\sum_{l_i=j}(\boldsymbol{x}_i-\boldsymbol{u}_j)(\boldsymbol{x}_i-\boldsymbol{u}_j)^\top, \\
\boldsymbol{S}_b &= \frac{1}{n}\sum_{j=1}^{c} n_j\boldsymbol{u}_j\boldsymbol{u}_j^\top, \\
\boldsymbol{S}_t &= \boldsymbol{S}_w+\boldsymbol{S}_b = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top,
\end{aligned} \tag{A.1}$$

where $\boldsymbol{S}_w$, $\boldsymbol{S}_b$, and $\boldsymbol{S}_t$ denote *within-class*, *between-class* and *total* scatter matrices respectively, $\boldsymbol{u}_j$ the class-wise centroids, and n_j the sample size of the j-th class (or person). The objective function of FDA aims at maximising $trace(\boldsymbol{S}_b)$ and minimising $trace(\boldsymbol{S}_w)$ simultaneously, where $\boldsymbol{S}_w$ can be replaced by $\boldsymbol{S}_t$ since $\boldsymbol{S}_t = \boldsymbol{S}_b + \boldsymbol{S}_w$. Hence, an optimal transformation $\boldsymbol{G}^*$ by FDA can be computed by solving the following problem:

$$\boldsymbol{G}^* = \arg\max_{\boldsymbol{G}} \; trace\Big(\big(\boldsymbol{G}^\top\boldsymbol{S}_b\boldsymbol{G}\big)\big(\boldsymbol{G}^\top\boldsymbol{S}_t\boldsymbol{G}\big)^\dagger\Big). \tag{A.2}$$

Theorem 1. *With* $\boldsymbol{Y}$ *defined as Eq.* (5.4)*, the projection* $\boldsymbol{P}^*$ *learned by Eq.* (5.3) *is equivalent to* $\boldsymbol{G}^*$*, the optimal FDA solution in Eq.* (A.2).

Proof. First, optimising the objective in Eq. (5.4) involves solving the following eigen-problem:

$$\boldsymbol{S}_t^{\dagger}\boldsymbol{S}_b\boldsymbol{G} = \boldsymbol{G}\boldsymbol{\Lambda}, \tag{A.3}$$

where $\boldsymbol{G} \in \mathbb{R}^{d\times q} = \left[\boldsymbol{g}_1, \cdots, \boldsymbol{g}_q\right]$ contains q eigenvectors of $\boldsymbol{S}_t^{\dagger}\boldsymbol{S}_b$, and $\boldsymbol{\Lambda} = diag(\alpha_1, \cdots, \alpha_q)$ with α_i the corresponding eigenvalue, and $q = rank(\boldsymbol{S}_b) \leq c-1$. From the definitions in Eq. (A.1) and Eq. (5.4), $\boldsymbol{S}_t$ and $\boldsymbol{S}_b$ can be further expanded as:

$$\boldsymbol{S}_t = \boldsymbol{X}\boldsymbol{X}^{\top}, \quad \boldsymbol{S}_b = \boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y}^{\top}\boldsymbol{X}^{\top}. \tag{A.4}$$

Here, the multiplier $\frac{1}{n}$ is omitted in both scatter matrices for simplicity. Now, we can rewrite the left-hand side of Eq. (A.3) as:

$$(\boldsymbol{X}\boldsymbol{X}^{\top} + \lambda\boldsymbol{I})^{\dagger}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y}^{\top}\boldsymbol{X}^{\top}\boldsymbol{G} = \boldsymbol{G}\boldsymbol{\Lambda}. \tag{A.5}$$

Note that, the pseudo-inverse $\boldsymbol{S}_t^{\dagger}$ is calculated by $(\boldsymbol{X}\boldsymbol{X}^{\top} + \lambda\boldsymbol{I})^{\dagger}$. The reason is that in real-world problems such as person re-id where training data is often less sufficient, $\boldsymbol{S}_t$ is likely to be ill-conditioned, i.e. singular or close to singular, so that its inverse cannot be accurately computed.

By our solution $\boldsymbol{P}$ in Eq. (5.3), we can further rewrite Eq. (A.5):

$$\boldsymbol{P}\boldsymbol{Y}^{\top}\boldsymbol{X}^{\top}\boldsymbol{G} = \boldsymbol{G}\boldsymbol{\Lambda} \tag{A.6}$$

To connect the regression solution $\boldsymbol{P}$ and the FDA solution $\boldsymbol{G}$, we define a $c \times c$ matrix $\boldsymbol{R} = \boldsymbol{Y}^{\top}\boldsymbol{X}^{\top}\boldsymbol{P}$. According to the general property of eigenvalues [170], $\boldsymbol{R}$ and $\boldsymbol{P}\boldsymbol{Y}^{\top}\boldsymbol{X}^{\top}$ share the same q non-zero eigenvalues. Also, if $\boldsymbol{V} \in \mathbb{R}^{c\times q}$ contains the q eigenvectors of $\boldsymbol{R}$, columns of the matrix $\boldsymbol{P}\boldsymbol{V}$ must be the eigenvectors of the matrix $\boldsymbol{P}\boldsymbol{Y}^{\top}\boldsymbol{X}^{\top}$. Therefore, the relation between $\boldsymbol{P}$ and $\boldsymbol{G}$ is:

$$\boldsymbol{G} = \boldsymbol{P}\boldsymbol{V} \tag{A.7}$$

Finally, we show in the following Lemma that $\boldsymbol{P}$ and $\boldsymbol{G}$ are equivalent in the aspect of re-id matching.

Lemma 1. *In the embedding provided by* $\boldsymbol{P}$ *and* $\boldsymbol{G}$*, the nearest neighbour algorithm produce same result. That is,* $(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\top}\boldsymbol{P}\boldsymbol{P}^{\top}(\boldsymbol{x}_i - \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\top}\boldsymbol{G}\boldsymbol{G}^{\top}(\boldsymbol{x}_i - \boldsymbol{x}_j)$.

Proof. The necessary and sufficient condition for Lemma 1 is $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{G}\boldsymbol{G}^\top$. As $\boldsymbol{V} \in \mathbb{R}^{c\times q}$, there must exist a matrix $\boldsymbol{V}_2 \in \mathbb{R}^{c\times(c-q)}$ such that $\hat{\boldsymbol{V}} = [\boldsymbol{V}, \boldsymbol{V}_2]$ is a $c \times c$ orthogonal matrix. Suppose the diagonal matrix $\boldsymbol{\Gamma}$ contains the non-zero eigenvalues of $\boldsymbol{R}$, then the eigen decomposition $\boldsymbol{R} = \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{V}^\top$ implies that $\boldsymbol{V}_2^\top \boldsymbol{R}\boldsymbol{V}_2 = 0$.

Recall that $\boldsymbol{R} = \boldsymbol{Y}^\top\boldsymbol{X}^\top\boldsymbol{P}$, and $\boldsymbol{P} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I})^\dagger\boldsymbol{X}\boldsymbol{Y}$, then we obtain:

$$\boldsymbol{V}_2^\top\boldsymbol{Y}^\top\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I})^\dagger\boldsymbol{X}\boldsymbol{Y}\boldsymbol{V}_2 = 0 \tag{A.8}$$

As $(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I})^\dagger$ is positive definite, the above equation implies that $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{V}_2 = 0$, and hence $\boldsymbol{P}\boldsymbol{V}_2 = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I})^\dagger\boldsymbol{X}\boldsymbol{Y}\boldsymbol{V}_2 = 0$. Hence, we have:

$$\begin{aligned} \boldsymbol{P}\boldsymbol{P}^\top &= \boldsymbol{P}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^\top\boldsymbol{P}^\top \\ &= \boldsymbol{P}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{P}^\top + \boldsymbol{P}\boldsymbol{V}_2\boldsymbol{V}_2^\top\boldsymbol{P}^\top \\ &= \boldsymbol{G}\boldsymbol{G}^\top + 0 \end{aligned} \tag{A.9}$$

As such, the proof to Lemma 1 and Theorem 1 is complete.

Bibliography

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, 2014.

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, pp. 1116–1124, 2015.

[3] S. Gong, M. Cristani, C. L. Chen, and T. M. Hospedales, "The re-identification challenge," in *Person Re-Identification*, pp. 1–20, Springer, 2014.

[4] "Closed-Circuit Television." `https://en.wikipedia.org/wiki/Closed-circuit_television`. Wikipedia Online, accessed 3 June 2017.

[5] "One surveillance camera for every 11 people in Britain, says CCTV survey." `http://www.telegraph.co.uk/technology/10172298`. The Telegraph, accessed 3 June 2017.

[6] "Video surveillance market by system (analog, ip, biometrics), hardware (camera, monitors, servers, storage devices), software (video analytics, vms), and service (vsaas, installation and maintenance), vertical, and region - global forecast to 2022." `http://www.marketsandmarkets.com/PressReleases/global-video-surveillance-market.asp`. Markets and Markets, accessed 3 June 2017.

[7] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition and tracking," in *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, vol. 3, pp. 1–7, Citeseer, 2007.

[8] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *British Machine Vision Conference*, vol. 2, 2009.

[9] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1258, 2016.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

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

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*, pp. 21–37, Springer, 2016.

[14] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *IEEE International Conference on Computer Vision*, pp. 1377–1385, 2015.

[15] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

[16] H. Keval and M. A. Sasse, "Not the Usual Suspects: A study of factors reducing the effectiveness of CCTV," *Security Journal*, vol. 23, no. 2, pp. 134–154, 2010.

[17] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

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

[19] W.-S. Zheng, S. Gong, and T. Xiang, “Re-identification by relative distance comparison,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 653–668, March 2013.

[20] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian, “Local fisher discriminant analysis for pedestrian re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3318–3325, 2013.

[21] R. Zhao, W. Ouyang, and X. Wang, “Learning mid-level filters for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[22] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by discriminative selection in video ranking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January 2016.

[23] F. Xiong, M. Gou, O. Camps, and M. Sznaier, “Person re-identification using kernel-based metric learning methods,” in *European Conference on Computer Vision*, pp. 1–16, Springer, 2014.

[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206, 2015.

[25] S. Liao and S. Z. Li, “Efficient PSD constrained asymmetric metric learning for person re-identification,” in *IEEE International Conference on Computer Vision*, pp. 3685–3693, 2015.

[26] “London stations continue to dominate ’top of the stops’ charts.” `http://orr.gov.uk/news-and-media/press-releases/2015/london-stations-continue-to-dominate-top-of-the-stops-charts`. Office of Rail and Road (ORR), accessed 15 December 2015.

[27] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, January 2014.

[28] P. Dollr, S. Belongie, and P. Perona, “The fastest pedestrian detector in the west,” in *British Machine Vision Conference*, vol. 2, p. 7, 2010.

[29] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling.," in *IEEE International Conference on Computer Vision*, pp. 32–39, 2009.

[30] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in *IEEE International Workshop on Visual Surveillance*, pp. 3–10, 2000.

[31] Y. Raja and S. Gong, "Scalable multi-camera tracking in a metropolis," in *Person Re-Identification*, pp. 413–438, Springer, 2014.

[32] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

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

[34] M. Hirzer, P. M. Roth, M. Koestinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification.," in *European Conference on Computer Vision*, pp. 780–793, 2012.

[35] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, pp. 262–275, Springer, 2008.

[36] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking.," in *British Machine Vision Conference*, vol. 2, p. 6, 2010.

[37] C. Liu, S. Gong, and C. L. Chen, "On-the-fly feature importance mining for person re-identification," *Pattern Recognition*, vol. 47, no. 4, pp. 1602–1615, 2014.

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

[39] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference*, vol. 2, p. 6, 2011.

[40] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human re-identification by matching compositional template with cluster sampling," in *IEEE International Conference on Computer Vision*, pp. 3152–3159, 2013.

[41] R. Layne, T. Hospedales, and S. Gong, "Person re-identification by attributes," in *British Machine Vision Conference*, vol. 2, p. 8, 2012.

[42] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes.," in *Workshop of European Conference on Computer Vision*, pp. 402–412, Springer, 2012.

[43] A. Li, L. Liu, and S. Yan, "Person re-identification by attribute-assisted clothes appearance," in *Person Re-Identification*, pp. 119–138, Springer, 2014.

[44] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Workshop of European Conference on Computer Vision*, pp. 413–422, Springer, 2012.

[45] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision*, pp. 536–551, Springer, 2014.

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

[47] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1239–1248, 2016.

[48] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *European Conference on Computer Vision*, pp. 405–422, Springer, 2016.

[49] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1278–1287, 2016.

[50] E. Ahmed, M. J. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3908–3916, 2015.

[51] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *European Conference on Computer Vision*, pp. 732–748, Springer, 2016.

[52] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*, pp. 791–808, Springer, 2016.

[53] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison.," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 649–656, 2011.

[54] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, pp. 688–703, Springer, 2014.

[55] J. Chen, Z. Zhang, and Y. Wang, "Relevance metric learning for person re-identification by exploiting listwise similarities," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4741–4755, 2015.

[56] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[57] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, 2016.

[58] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1288–1296, 2016.

[59] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2666–2672, 2012.

[60] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pp. 244–249, IEEE, 2013.

[61] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *ACM International Conference on Distributed Smart Cameras*, p. 10, 2014.

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

[63] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*, pp. 135–153, Springer, 2016.

[64] W.-S. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification.," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2650–2657, 2012.

[65] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *IEEE International Conference on Computer Vision*, pp. 3567–3574, 2013.

[66] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *Workshop of European Conference on Computer Vision*, pp. 381–390, Springer, 2012.

[67] R. Layne, T. M. Hospedales, and S. Gong, "Domain transfer for person re-identification," in *Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, pp. 25–32, 2013.

[68] B. Ma, Y. Su, and F. Jurie, "BiCov: a novel image representation for person re-identification and face verification," in *British Machine Vision Conference*, p. 11, 2012.

[69] C. Liu, C. C. Loy, S. Gong, and G. Wang, "POP: Person re-identification post-rank optimisation," in *IEEE International Conference on Computer Vision*, pp. 441–448, 2013.

[70] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Columbus, Ohio, United States), June 2014.

[71] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, pp. 1–114, December 2012.

[72] J. Kang, K. R. Ryu, and H.-C. Kwon, "Using cluster-based sampling to select initial training set for active learning in text classification," in *Advances in knowledge discovery and data mining*, pp. 384–388, Sydney, Australia: Springer, May 2004.

[73] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[74] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12, Springer-Verlag New York, Inc., 1994.

[75] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *the eleventh international conference on machine learning*, pp. 148–156, 1994.

[76] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2-3, pp. 133–168, 1997.

[77] N. Roy and A. McCallum, "Toward optimal active learning through monte carlo estimation of error reduction," in *International Conference on Machine learning*, pp. 441–448, 2001.

[78] T. Osugi, D. Kim, and S. Scott, "Balancing exploration and exploitation: A new algorithm for active machine learning," in *IEEE International Conference on Data Mining*, (Houston, Texas, United States), pp. 8–pp, November 2005.

[79] N. Cebron and M. R. Berthold, "Active learning for object classification: From exploration to exploitation," *Data Mining and Knowledge Discovery*, vol. 18, no. 2, pp. 283–299, 2009.

[80] T. M. Hospedales, S. Gong, and T. Xiang, "A unifying theory of active discovery and learning," in *European Conference on Computer Vision*, pp. 453–466, Springer, 2012.

[81] S. Ebert, M. Fritz, and B. Schiele, "Ralf: A reinforced active learning formulation for object class recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Providence, Rhode Island, United States), pp. 3626–3633, June 2012.

[82] C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong, "Stream-based joint exploration-exploitation active learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Providence, Rhode Island, United States), pp. 1560–1567, June 2012.

[83] C. Käding, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler, "Active learning and discovery of object categories in the presence of unnameable instances," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4343–4352, 2015.

[84] Z. Wang, B. Du, L. Zhang, L. Zhang, M. Fang, and D. Tao, "Multi-label active learning based on maximum correntropy criterion: Towards robust and discriminative labeling," in *European Conference on Computer Vision*, pp. 453–468, Springer, 2016.

[85] A. Das, R. Panda, and A. Roy-Chowdhury, "Active image pair selection for continuous person re-identification," in *IEEE International Conference on Image Processing*, pp. 4263–4267, 2015.

[86] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *European Conference on Computer Vision*, pp. 858–877, 2016.

[87] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Advances in neural information processing systems*, pp. 409–415, 2001.

[88] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 31, pp. 497–508, 2001.

[89] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: fast feature extraction and svm training," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1689–1696, 2011.

[90] J. Sánchez and F. Perronnin, "High-dimensional signature compression for large-scale image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1665–1672, 2011.

[91] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, “Distance-based image classification: Generalizing to new classes at near-zero cost,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2624–2637, 2013.

[92] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, “Incremental learning of ncm forests for large-scale image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3654–3661, 2014.

[93] A. Yao, J. Gall, C. Leistner, and L. Van Gool, “Interactive object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3242–3249, 2012.

[94] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, “Incremental learning for robust visual tracking,” *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[95] H. Wang, S. Gong, and T. Xiang, “Highly efficient regression for scalable person re-identification,” in *British Machine Vision Conference*, 2016.

[96] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, “Hierarchical gaussian descriptor for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1372, 2016.

[97] D. Singaraju, L. Grady, and R. Vidal, “Interactive image segmentation via minimization of quadratic energies on directed graphs,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[98] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” in *ACM Transactions on Graphics*, vol. 23, pp. 309–314, 2004.

[99] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, “Visual recognition with humans in the loop,” in *European Conference on Computer Vision*, pp. 438–451, Springer, 2010.

[100] C. Wah, S. Branson, P. Perona, and S. Belongie, “Multiclass recognition and part localization with humans in the loop,” in *IEEE International Conference on Computer Vision*, pp. 2524–2531, 2011.

[101] S. Lad and D. Parikh, “Interactively guiding semi-supervised clustering via attribute-based explanations,” in *European Conference on Computer Vision*, pp. 333–349, Springer, 2014.

[102] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, “Interactive object counting,” in *European Conference on Computer Vision*, pp. 504–518, Springer, 2014.

[103] X. S. Zhou and T. S. Huang, “Relevance feedback in image retrieval: A comprehensive review,” *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

[104] W.-C. Lin, Z.-Y. Chen, S.-W. Ke, C.-F. Tsai, and W.-Y. Lin, “The effect of low-level image features on pseudo relevance feedback,” *Neurocomputing*, 2015.

[105] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, and J. Luo, “Efficient manifold ranking for image retrieval,” in *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 525–534, 2011.

[106] B. Settles, “Active learning literature survey,” *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[107] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Scandinavian Conference on Image Analysis*, pp. 91–102, Springer, 2011.

[108] W. Li and X. Wang, “Locally aligned feature transforms across views.,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3594–3601, 2013.

[109] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *IEEE International Conference on Computer Vision*, pp. 2960–2967, 2013.

[110] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, November 2006.

[111] M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton, “Semi-supervised kernel canonical correlation analysis with application to human fmri,” *Pattern Recognition Letters*, vol. 32, no. 11, pp. 1572–1583, 2011.

[112] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[113] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pp. 2037–2041, 2006.

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

[115] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by salience matching,” in *IEEE International Conference on Computer Vision*, pp. 2528–2535, December 2013.

[116] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, “Pedestrian recognition with a learned metric,” in *Asian Conference on Computer Vision*, pp. 501–512, Springer, 2010.

[117] Z. Shi, T. M. Hospedales, and T. Xiang, “Bayesian joint topic modelling for weakly supervised object localisation,” in *IEEE International Conference on Computer Vision*, pp. 2984–2991, 2013.

[118] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[119] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[120] Z. Zhang, G. Dai, C. Xu, and M. I. Jordan, “Regularized discriminant analysis, ridge regression and beyond,” *The Journal of Machine Learning Research*, vol. 11, pp. 2199–2228, 2010.

[121] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[122] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 2013.

[123] F. Zhu, J. Xie, and Y. Fang, “Heat diffusion long-short term memory learning for 3d shape analysis,” in *European Conference on Computer Vision*, 2016.

[124] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu, “Random faces guided sparse many-to-one encoder for pose-invariant face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2416–2423, 2013.

[125] M. A. Woodbury, "Inverting modified matrices," *Memorandum report*, vol. 42, p. 106, 1950.

[126] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2372–2379, 2009.

[127] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the conference on empirical methods in natural language processing*, pp. 1070–1079, Association for Computational Linguistics, 2008.

[128] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, pp. 199–213, Springer, 1998.

[129] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1301–1306, 2010.

[130] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Journal of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[131] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1629–1642, 2015.

[132] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *IEEE International Conference on Image Processing*, pp. 769–773, 2016.

[133] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.

[134] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. C. Yuen, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1661–1675, 2017.

[135] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 1–8, 2016.

[136] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1268–1277, 2016.

[137] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, December 2009.

[138] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning.," in *International Conference on Machine learning*, pp. 209–216, ACM, 2007.

[139] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person re-identification with reference descriptor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 776–787, 2016.

[140] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised L1 graph learning," in *European Conference on Computer Vision*, pp. 178–195, Springer, 2016.

[141] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[142] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Advances in Neural Information Processing Systems*, pp. 4170–4178, 2016.

[143] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European Conference on Computer Vision*, pp. 475–491, Springer, 2016.

[144] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[145] R. Penrose, “A generalized inverse for matrices,” in *Proc. Cambridge Philos. Soc*, vol. 51, pp. 406–413, Cambridge Univ Press, 1955.

[146] D. Cai, X. He, and J. Han, “Srda: An efficient algorithm for large-scale discriminant analysis,” *Data Mining and Knowledge Discovery*, vol. 20, no. 1, pp. 1–12, 2008.

[147] R. E. Schapire, “The strength of weak learnability,” *Machine Learning*, vol. 5, pp. 197–227, July 1990.

[148] Y. Amit and D. Geman, “Shape quantization and recognition with randomized trees,” *Neural Computing*, vol. 9, pp. 1545–1588, October 1997.

[149] D. Parkhurst, K. Law, and E. Niebur, “Modeling the role of salience in the allocation of overt visual attention,” *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[150] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, “Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1153–1162, 2016.

[151] D. Lim and G. Lanckriet, “Efficient learning of mahalanobis metrics for ranking,” in *International Conference on Machine learning*, pp. 1980–1988, 2014.

[152] J. Weston, S. Bengio, and N. Usunier, “Large scale image annotation: learning to rank with joint word-image embeddings,” *Machine learning*, vol. 81, no. 1, pp. 21–35, 2010.

[153] N. Usunier, D. Buffoni, and P. Gallinari, “Ranking with ordered weighted pairwise classification,” in *International Conference on Machine learning*, pp. 1057–1064, ACM, 2009.

[154] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large scale online learning of image similarity through ranking,” *The Journal of Machine Learning Research*, pp. 1109–1135, March 2010.

[155] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of International Conference on Computational Statistics*, pp. 177–186, Springer, 2010.

[156] K. Tsuda, G. Rätsch, and M. K. Warmuth, “Matrix exponentiated gradient updates for on-line learning and Bregman projection,” in *Journal of Machine Learning Research*, pp. 995–1018, 2005.

[157] J. Kivinen and M. K. Warmuth, “Exponentiated gradient versus gradient descent for linear predictors,” *Information and Computation*, pp. 1–63, 1997.

[158] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *Proceedings of the 24th international conference on Machine learning*, pp. 209–216, ACM, 2007.

[159] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman, “Online metric learning and fast similarity search,” in *Advances in Neural Information Processing Systems*, pp. 761–768, 2009.

[160] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1.” `http://cvxr.com/cvx`, March 2014.

[161] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[162] M. Geng, Y. Wang, T. Xiang, and Y. Tian, “Deep transfer learning for person re-identification,” *arXiv preprint arXiv:1611.05244*, 2016.

[163] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “End-to-end deep learning for person search,” *arXiv preprint arXiv:1604.01850*, 2016.

[164] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys*, pp. 5:1–5:60, April 2008.

[165] R. Zhao, W. Ouyang, H. Li, and X. Wang, “Saliency detection by multi-context deep learning,” in *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265–1274, 2015.

[166] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.

[167] N. Xu, B. Price, S. Cohen, and T. Huang, “Deep image matting,” *arXiv preprint arXiv:1703.03872*, 2017.

[168] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[169] J. C. Caicedo and S. Lazebnik, “Active object localization with deep reinforcement learning,” in *The IEEE International Conference on Computer Vision*, pp. 2488–2496, 2015.

[170] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.