Cross-class Transfer Learning for Visual Data

Elyor Kodirov

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

School of Electronic Engineering and Computer Science

Queen Mary University of London

28 November 2017

Cross-class Transfer Learning for Visual Data

Elyor Kodirov

Abstract

Automatic analysis of visual data is a key objective of computer vision research; and performing visual recognition of objects from images is one of the most important steps towards understanding and gaining insights into the visual data. Most existing approaches in the literature for the visual recognition are based on a supervised learning paradigm. Unfortunately, they require a large amount of labelled training data which severely limits their *scalability*. On the other hand, recognition is instantaneous and effortless for humans. They can recognise a new object without seeing any visual samples by just knowing the description of it, leveraging similarities between the description of the new object and previously learned concepts. Motivated by humans recognition ability, this thesis proposes novel approaches to tackle *cross-class transfer learning* (cross-class recognition) problem whose goal is to learn a model from seen classes (those with labelled training samples) that can generalise to unseen classes (those with labelled testing samples) without any training data i.e., seen and unseen classes are disjoint. Specifically, the thesis studies and develops new methods for addressing three variants of the cross-class transfer learning:

Chapter 3 The first variant is *transductive cross-class transfer learning*, meaning labelled training set and unlabelled test set are available for model learning. Considering training set as the source domain and test set as the target domain, a typical cross-class transfer learning assumes that the source and target domains share a common *semantic space*, where visual feature vector extracted from an image can be embedded using an embedding function. Existing approaches learn this function from the source domain and apply it without adaptation to the target one. They are therefore prone to the domain shift problem i.e., the embedding function is only concerned with predicting the training seen class semantic representation in the learning stage during learning, when applied to the test data it may underperform. In this thesis, a novel cross-class transfer learning (CCTL) method is proposed based on unsupervised domain adaptation. Specifically, a novel regularised dictionary learning framework is formulated by which the target class labels are used to regularise the learned target domain embeddings thus effectively overcoming the projection domain shift problem.

Chapter 4 The second variant is *inductive cross-class transfer learning*, that is, only training set is assumed to be available during model learning, resulting in a harder challenge compared to the previous one. Nevertheless, this setting reflects a real-world setting in which test data is available after the model learning. The main problem remains the same as the previous variant, that is, the domain shift problem occurs when the model learned only from the training set is applied to the test set without adaptation. In this thesis, a semantic autoencoder (SAE) is proposed building on an encoder-decoder paradigm. Specifically, first a semantic space is defined so that knowledge transfer is possible from the seen classes to the unseen classes. Then, an encoder aims to embed/project a visual feature vector into the semantic space. However, the decoder exerts a generative task, that is, the projection must be able to reconstruct the original visual features. The

generative task forces the encoder to preserve richer information, thus the learned encoder from seen classes is able generalise better to the new unseen classes.

Chapter 5 The third one is *unsupervised cross-class transfer learning*. In this variant, no supervision is available for model learning i.e., only unlabelled training data is available, leading to the hardest setting compared to the previous cases. The goal, however, is the same, learning some knowledge from the training data that can be transferred to the test data composed of completely different labels from that of training data. The thesis proposes a novel approach which requires no labelled training data yet is able to capture discriminative information. The proposed model is based on a new graph regularised dictionary learning algorithm. By introducing a l_1-norm graph regularisation term, instead of the conventional squared l_2-norm, the model is robust against outliers and noises typical in visual data. Importantly, the graph and representation are learned jointly, resulting in further alleviation of the effects of data outliers. As an application, person re-identification is considered for this variant 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 works have been published at peer-reviewed conferences:

Chapter 3

1. E. Kodirov, T. Xiang, Z. Fu and S. Gong. *Unsupervised Domain Adaptation for Zero-shot Learning*. In Proc. IEEE Conference on Computer Vision, Santiago, Chile, 2015. **(ICCV)**

Chapter 4

1. E. Kodirov, T. Xiang and S. Gong. *Semantic Autoencoder for Zero-shot Learning*. In Proc. IEEE International Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 2017. Spotlight, **(CVPR)**

Chapter 5

1. E. Kodirov, T. Xiang, Z. Fu and S. Gong. *Person Re-identification by Unsupervised L_1 Graph Learning*. International European Conference on Computer Vision, Amsterdam, Netherland, 2016. **(ECCV)**
2. E. Kodirov, T. Xiang, Z. Fu and S. Gong. *Learning Robust Graph Regularisation for Subspace Clustering*. British Machine Vision Conference, York, UK, 2016. Oral, **(BMVC)**
3. E. Kodirov, T. Xiang, Z. Fu and S. Gong. *Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification*. British Machine Vision Conference, Swansea, UK, 2015. **(BMVC)**

Acknowledgements

Firstly, I would like to express my deep gratitude to my respected supervisor Prof. Tao Xiang for his perpetual patience and enthusiastic supervision. He is always there whenever I need help, and I am very grateful spending my time, about 3.5 years, with him. Also, I convey my special thanks to Dr. Zhenyong Fu and second-supervisor Prof. Shaogang Gong for their guidance and productive support. Owing to their excellent supervision, I gradually learn to derive new research ideas, and conduct independent research. I would like to thank my independent assessor Dr. Ioannis Patras for his support during my PhD study.

I am also very grateful with all friendly and supportive VISION GROUP members and visitors from different places: Tim Hospedales, Yi-Zhe Song, Miles Hansard, Xiatian Zhu, Li Zhang, Tanmoy Mukherjee, Zhiyuan Shi, Yanwei Fu, Ryan Layne, Feng Liu, Yi Li, Xun Xu, Hanxiao Wang, Yongxin Yang, Ioannis Alexiou, Jingya Wang, Qian Yu, Kunkun Pang, Yaowei Wang, Shuxin Ouyang, Xiangyu Kong. I am really pleased from QMUL administrative and system support staff for their great assistance.

Definitely, I am deeply thankful to my family members. They in particular my parents, my sisters, my brothers provide motivational ‘energy’ throughout my PhD. Importantly, I have to thank my brother Eldor Ibragimov and my loving wife Khusinobod Turabova for their devotion all the time.

Contents

1 Introduction 14

1.1 Scope of the Thesis 14

1.2 Problem Definition 17

1.2.1 Cross-class Transfer Learning 17

1.2.2 Variants of Cross-class Transfer Learning 18

1.3 Challenges and Solutions 18

1.3.1 Transductive Cross-class Transfer Learning 20

1.3.2 Inductive Cross-class Transfer Learning 22

1.3.3 Unsupervised Cross-class Transfer Learning 22

1.4 Contributions 24

1.5 Thesis Outline 24

2 Literature Review 27

2.1 Machine Learning Tools and Visual Recognition: An Overview 27

2.1.1 Machine Learning 27

2.1.2 Transfer Learning 29

2.1.3 Visual Recognition 32

2.2 Transductive Cross-class Transfer Learning 35

2.3 Inductive Cross-class Transfer Learning 43

2.4 Unsupervised Cross-class Transfer Learning 45

2.5 Benchmark Datasets 51

3 Transductive Cross-class Transfer Learning by Unsupervised Domain Adaptation 54

3.1 Methodology 55

3.1.1 Problem Formulation 55

3.1.2 Dictionary Learning for Embedding Function Learning 56

3.1.3 Domain Adaptation by Regularised Dictionary Learning 58

3.1.4 Optimisation 59

3.1.5 Cross-class Recognition 59

3.2 Experiments and Evaluations 60

3.2.1 Datasets and Settings 60

3.2.2 Comparative evaluation 61

3.2.3 Comparison with Unsupervised Domain Adaptation Methods 66

3.2.4 Further Analysis 68

3.3 Conclusion 71

4 Inductive Cross-class Transfer Learning by Semantic Autoencoder 74

4.1 Methodology 76

4.1.1 Linear Autoencoder 76

4.1.2 Semantic Autoencoder 76

4.1.3 Optimisation 77

4.1.4 SEA for Cross-class Recognition 78

4.1.5 Relations to Existing Models 79

4.1.6 Deep SAE 80

4.2 Experiments and Evaluations 81

4.2.1 Datasets and Settings 81

4.2.2 Comparative evaluation 85

4.2.3 Ablation Study 86

4.2.4 DSAE results 86

4.2.5 Further Analysis 87

4.3 Conclusion 88

5 Unsupervised Cross-class Transfer Learning by L_1 Graph Learning 90

5.1 Methodology 93

5.1.1 Problem Formulation 93

5.1.2 Robust Graph Regularised Dictionary Learning 93

5.1.3 Optimisation 96

5.1.4 Bayesian Interpretation 100

5.1.5 Cross-view Matching 101

5.1.6 Extension to Supervised ReID 101

5.2 Experiments and Evaluations 102

5.2.1 Datasets and Settings 102

5.2.2 Evaluation of Unsupervised Learning based ReID 104

5.2.3 Evaluation of Supervised Learning based ReID 106

5.2.4 Further Analysis 107

5.3 Conclusion 110

6 Conclusion and Future Work **112**

6.1 Conclusion 112

6.2 Future Work 113

Appendices **115**

A Bartels-Stewart Algorithm **116**

B Label Propagation **117**

C Indirect Attribute Prediction **119**

List of Figures

1.1 The key challenges in the visual recognition. 15

1.2 Conventional supervised learning for visual recognition. 17

1.3 Three variants of cross-class transfer learning are studied. 19

1.4 Projection domain shift problem. 21

1.5 Person re-identification challenges in the wild. 23

1.6 A summary of main chapters and structure of all chapters. 26

2.1 A conventional machine learning setting. 30

2.2 An illustration of transfer learning setting. 30

2.3 An illustration of unsupervised domain adaptation setting. 31

2.4 Pipeline for category recognition (left), and instance verification (right). 33

2.5 Pipeline for HOG feature extraction. 35

2.6 Deep learning-based visual recognition. 36

2.7 An example of a deep learning model: VGG network. 36

2.8 An illustration of dictionary learning. 37

2.9 Label embedding. 38

2.10 Semantic space. 40

2.11 Cross-class transfer learning for category recognition. 41

2.12 Domain shift problem. 42

2.13 A linear autoencoder. 44

2.14 Different ways of embedding function learning. 45

2.15 Person re-identification. 48

2.16 Examples: AwA, CUB, aP&Y, and SUN datasets. 52

2.17 Examples from UCF101 dataset. 52

2.18 Examples from ReID datasets. 53

3.1 Evaluation of the contributions of each component of the DTCCTL. 68

3.2 Effectiveness of combining multiple semantic embedding spaces. 69

3.3 The effect of the amount of target data used. 70

3.4 Semantic dictionary visualisation. 72

3.5 Qualitative results on UCF101 dataset. 73

4.1 The proposed semantic autoencoder (SAE). 75

4.2 Different ways of learning embedding space. 80

4.3 Deep semantic autoencoder. 81

5.1 A graph constructed in the hand-crafted feature space vs. a graph learned using the proposed model in this work. 91

5.2 CMC curves for VIPeR and CUHK01. The curves correspond to Table 5.1. . . . 105

5.3 (Left) Rank 1 accuracies with different dictionary sizes on VIPeR dataset; (Right) Objective function value with respect to the number of iterations on CUHK01. . . 109

5.4 Dictionary visualisation. 110

5.5 An illustration of the graph **W** before and after learning process. 111

C.1 Indirect attribute prediction (IAP) 119

List of Tables

2.1 Benchmark datasets for evaluation of transductive and inductive cross-class transfer learning. 53

2.2 Benchmark datasets for evaluation of person re-identification. 53

3.1 Transductive cross-class recognition results on AwA and CUB on the target domain (%). 65

3.2 Transcductive cross-class recognition results on the UCF101 dataset. 66

3.3 Evaluations on unsupervised domain adaptation methods. 67

3.4 Classification methods: nearest neighbour (NN) vs. label propagation (LP). . . . 69

3.5 Evaluations on cross-dataset cross-class transfer learning. 70

3.6 Attribute prediction with and without domain adaptation. 71

4.1 Comparative cross-class recognition accuracy (%, hit@5 for large-scale datasets) 82

4.2 The importance of adding the reconstruction constraint 86

4.3 Deep SAE results on AwA and CUB. Accuracy and training time are reported. DSAE-3 - '3' refers to the number of mid-level layers. 87

4.4 Comparative evaluation on generalised cross-class transfer learning on AwA and CUB. 88

4.5 Evaluating the computational cost (in second) on AwA. Encoder is used. 88

5.1 Unsupervised ReID results measured in Rank-1 matching accuracy (%) on VIPeR, PRID, CUHK01, CUHK03. 106

5.2 Comparison with state-of-the-art supervised learning-based methods. 108

5.3 The contributions of individual model components of the L_1Graph. 109

5.4 Average testing time of different methods on VIPeR. 109

General rules for notation definition

scalar	normal lower-case letters
set	normal UPPER-case letters
vector	**bold** lower-case letters
matrix	**bold** UPPER-case letters

Abbreviations

CCTL	Cross-class transfer learning
TCCTL	Transductive cross-class transfer learning
UCCTL	Unsupervised cross-class transfer learning
ILSVRC	ImageNet large scale visual recognition challenge
SAE	Semantic autoencoder
ZSL	Zero-shot learning
ReID	Person re-identification
CMC	Cumulative matching curve
UDA	Unsupervised domain adaptation
DNN	Deep neural networks
ANN	Artificial neural networks
CNN	Convolutional neural networks
PUL	Progressive unsupervised learning
AR	Adaptation regularisation constraint
VSS	Visual-semantic similarity constraint
DTCCTL	Dictionary learning-based cross-class transfer learning
NN	nearest neighbour
LP	label propagation
AwA	Animal with attributes
CUB	Caltech-UCSD Birds
L	hand-crafted features
MBH	Motion boundary histogram
ADMM	Alternating direction method of multipliers
AUSUC	Area under seen-unseen accuracy curve
HOG	Histogram of gradients
SIFT	Scale-invariant feature transform
LBP	Local binary pattern
GFK	Geodesic flow kernel
SADA	Subspace alignment domain adaptation
SIDL	Subspace interpolation dictionary learning
FV	Fisher vector
A	Attribute space
W	Word vector space
H	Word hierarchy
SI	Side information
CS	Class similarity
AE	Autoencoder
KNN	k-nearest neighbour
DSAE	Deep semantic autoencoder

Chapter 1

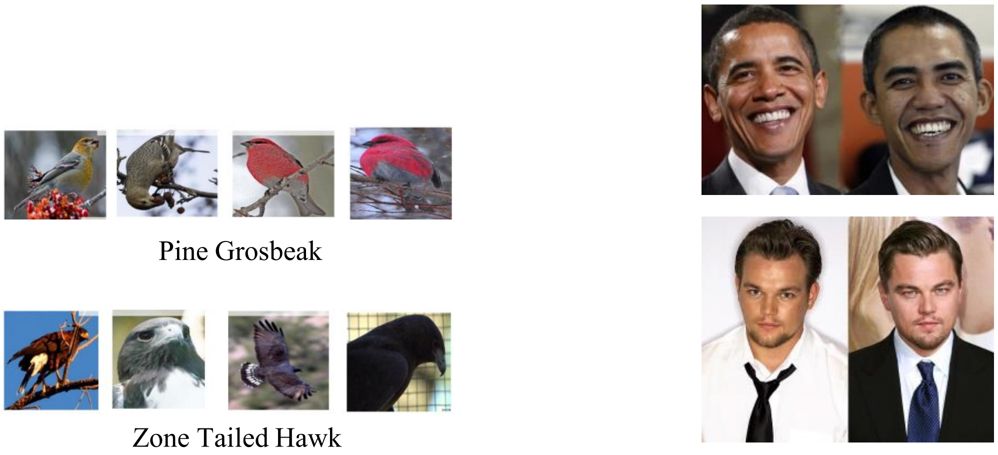
Introduction

1.1 Scope of the Thesis

Being able to understand the content of an image via automated visual analysis is one of the ultimate goals of computer vision research. The analysis could take many different forms depending on the task that is to be performed. Among them, visual recognition, which is a main topic in this thesis, is arguably the most important step in visual data analysis in which it allows to understand and gain insights into the visual data e.g., recognising objects, scenes, and categories. It can benefit tremendously a range of real-world applications that touch upon many areas of artificial intelligence and information retrieval, for example, surveillance systems, content-based image search, self-driving cars, or object identification for mobile robots.

There are in general two types of visual recognition tasks: (1) category recognition, and (2) instance recognition. In the category case, one seeks to recognise different instances of a generic category found in the visual data (images and videos) as belonging to same conceptual class such as person, animal, or car. In contrast, the goal of instance recognition is to recognise instances of a specific object or scene e.g., retrieving a specific person's image from gallery images for the purpose of verification.

Nevertheless, visual recognition is challenging. The key challenge is how to cope with the large intra-class variation and small inter-class variation. As depicted in Figure 1.1: for the former, a single bird species can have large intra-class variation: pose variation, background variation and appearance variation (see Figure 1.1a) whilst, for the latter, two persons are very similar



(a) Large intra-class variation

(b) Small inter-class variation

Figure 1.1: The key challenges in the visual recognition. (a) Large intra-class variation: each row shows images from the same species. For each bird species there are large intra-class variations: pose variation, background variation and appearance variation. (b) Small inter-class variation: although people have different identities (classes), they are visually very similar.

visually, although they have different identities (see Figure 1.1b). This challenge becomes even harder when images are partially occluded, or composed of unrelated background 'clutter'. A great deal of effort have been taken by researchers to tackle this challenge. They mainly focus on learning discriminative features and/or classifier that can maximise the inter-class variation, and minimise the intra-class one. To accomplish that, one can follow a *supervised learning* paradigm (Kotsiantis, 2007): first, typically enough examples with corresponding labels/annotations are collected for training a model; labels are there to guide the learning process to ensure separability of the different classes. Then, the model is applied to the test data examples of the same classes (see Figure 1.2).

A recent endeavour in visual recognition research is to perform large-scale recognition i.e., recognising thousands of category types. Unfortunately, conventional recognition algorithms face the problem of *scalability*, meaning when conventional approaches are considered to large-scale recognition, they require collecting large quantities of annotated instances for each class; collecting unambiguously/high quality labelled image/video examples are prohibitively expensive. For instance, the popular ImageNet large scale visual recognition challenge (ILSVRC), mainly focuses on the task of recognising 1K classes, a rather small subset of the full ImageNet dataset consisting of more than 21K classes with 14M images (Krizhevsky et al., 2012). This is because many of the 21K classes are only composed of a handful of images including 296 classes with only one image. Furthermore, this is more problematic when visual recognition research moves towards a finer granularity. For example, naming many fine-grained bird classes (e.g., ‘Pine Grosbeak’) is very challenging for most people, let alone collecting instances. Moreover, they cannot deal with the case that new classes may appear after the learning stage e.g., considering an ever growing set of classes, such as detecting new species of living beings and designing new products.

On the other hand, humans has a remarkable ability to perform visual recognition effortlessly, and instantaneously. There are a great deal of psychological/biological research on how humans gain such ability. The researchers found that one of the main reasons is transfer learning in which knowledge gained from one domain is abstracted and reused in other domains. This principle is a central part of understanding how people learn in general. That is, the humans construct new knowledge by integrating new concepts and propositions with related concepts and propositions they already aware of Ausubel (1968). Concretely, they are great at recognising a new object without seeing any visual samples by just knowing the description of it, leveraging similarities between the description of the new object and previously learned concepts Romera-Paredes and Torr (2015). For example, a human would have no problem of recognising a ’yellra’ if he has seen zebra before and also learned that a ’yellra’ is like zebra but black-and-yellow strips on it. The humans can easily generalise the knowledge learned in the past to recognise the classes never seen before. This transfer learning behaviour of humans has been demonstrated in children as young as 3 years by Brown and Kane (1988). Recently Canini et al. (2010) studied how human learners exhibit transfer learning. To accomplish this, they proposed two laboratory experiments that measure the degree to which humans engage in transfer learning. They particularly focus on people who learn systems of inter-related categories consisting of shared clusters. They experimentally confirmed that human learners engage in transfer learning in categorisation tasks. To support the discussion above, Caas (2010) explains schematically the key memory systems in the human brain. According to him, the brain is not an ’empty vessel’ to be filled with information, rather it is an extremely complex system that integrates and stores sensory and semantic information. In particular, information in short-term and ’working’ memory interacts with knowledge in our long-term memory, in order to create new meanings and long-term memories.

Very recently, researchers in machine learning and computer vision community have started to propose approaches that imitate the humans’ recognition ability, and this is known as *cross-*

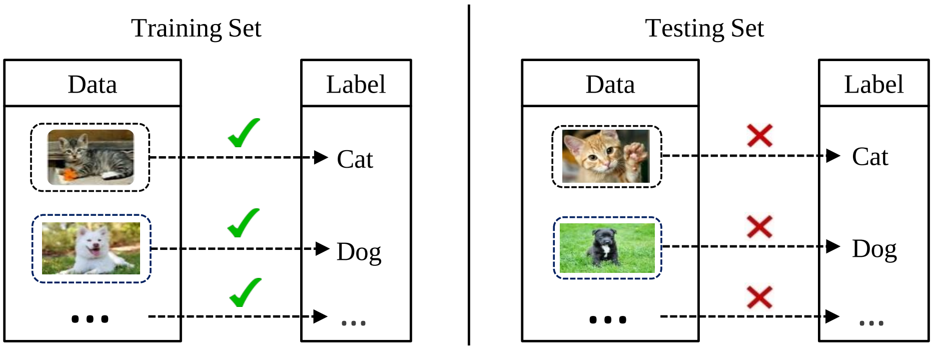


Figure 1.2: An illustration of conventional supervised learning for visual recognition. In the learning stage, training set is used to learn a visual recognition model, then that model is applied to perform recognition on a new test example, coming from the test set. The training set/testing set is composed of observed examples (samples) and corresponding manually annotated labels. The training set and test set belong to the same set of labels. That is, if training set contains images labelled as 'cat' or 'dog', then in the inference stage, test set also contains images labelled as either 'cat' or 'dog'. 'Yes' tick means that datum and corresponding label are given. 'No' tick for no correspondence between datum and label, and it needs to be found in the inference stage.

class transfer learning (CCTL) (Guo et al., 2016b). To this end, this thesis studies and develops novel approaches for addressing the cross-class transfer learning problem.

1.2 Problem Definition

1.2.1 Cross-class Transfer Learning

In cross-class transfer learning (CCTL), the class labels for the training set and test set are *disjoint*, unlike standard supervised learning methods as shown in Figure 1.2. In CCTL, however, there is an assumption that the training set and test set are related through a so called *semantic space*. This space plays the role of a bridge (ontology) that connects between the training data classes and the test data classes. The training data classes and the test data classes are often referred to as seen classes and unseen classes, respectively. Depending on the context the training set and test set are also called as source domain and target domain, respectively. A typical approach to CCTL requires to extract the knowledge from the labelled training data (seen classes) and to transfer that knowledge to the test set (unseen classes). This thesis focuses on cross-class transfer learning problem whose aim is to learn a transferable knowledge in the training stage that can generalise to test data (unseen classes) which is composed of completely different set of labels from the training set (seen classes). The name of 'cross-class' is coming from the fact that the training set and the test set are disjoint.

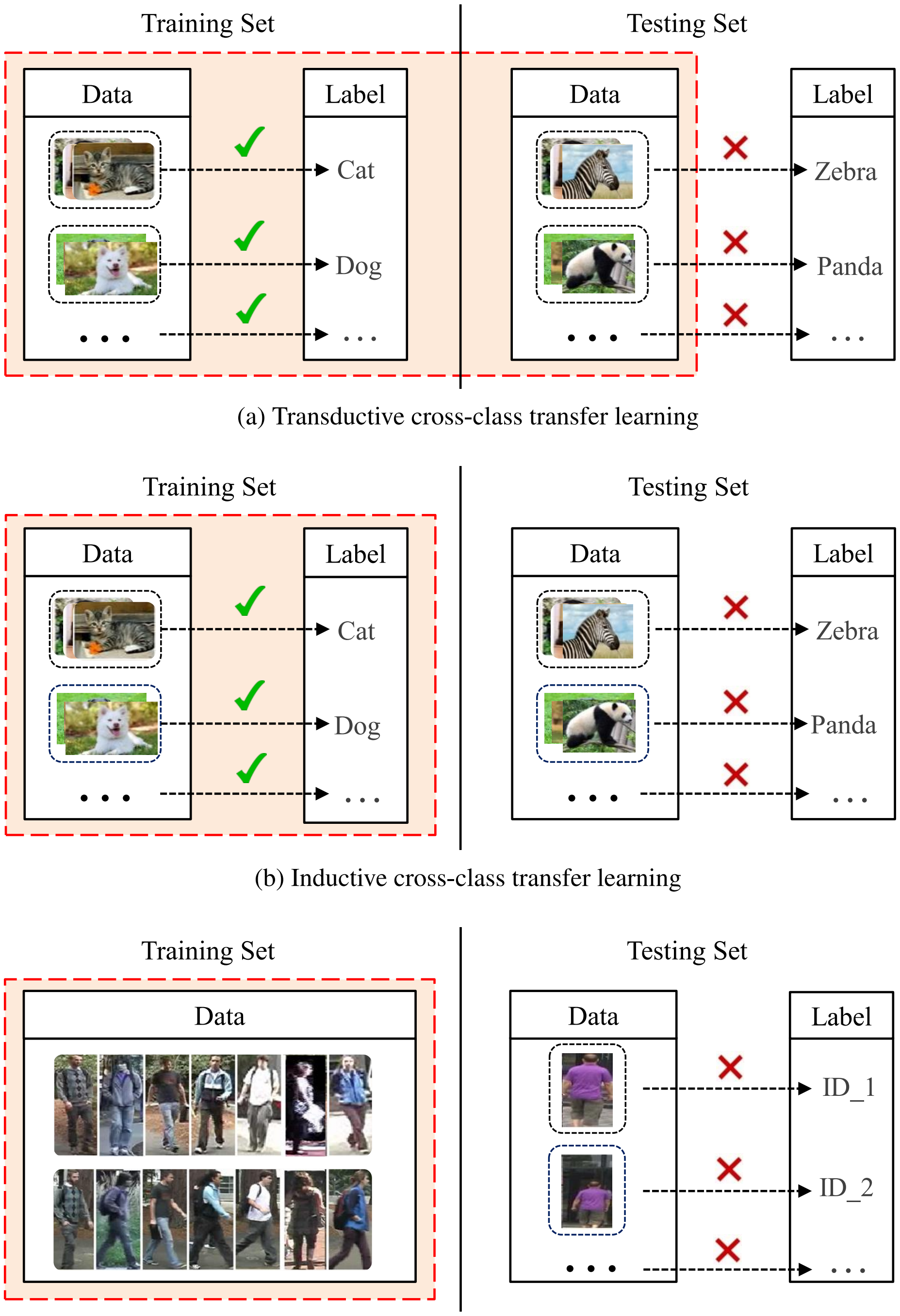
1.2.2 Variants of Cross-class Transfer Learning

There are different variants of the CCTL depending on the availability of supervision associated with the data i.e., labels. In this thesis, three variants of the CCTL ranging from 'easiest' to 'hardest' are considered, and they are illustrated in Figure 1.3. The first two variants deal with category recognition/classification problem, while the third one tackles instance-level verification problem. In all variants, training classes and test classes are disjoint, unlike the conventional transfer learning methods (Pan and Yang, 2010). Specifically,

1. *Transductive cross-class transfer learning*: The underlying aim is the category recognition. In this case, it is assumed that labelled training data and test set without corresponding labels are available (see Figure 1.3a). The unlabelled test data is used to reduce the distribution mismatch between training data and test data during a model learning so that the model can generalise to the test data well. Note that the full test set is used for model learning in this variant.
2. *Inductive cross-class transfer learning*: The underlying aim is the same as the previous variant. However, more practical setting of CCTL is considered in which it is not assumed that any unlabelled data from the test set is accessible (see Figure 1.3b). Zero-shot classification is considered as an application for this setting.
3. *Unsupervised cross-class transfer learning*: The goal is to learn a transferable knowledge that can be used for instance recognition such as face/person verification. In this variant, it is assumed that any means of labelled information associated with the data is unavailable in the learning stage unlike previous two variants. Yet, knowledge is learned by only relying on unlabelled training data in an unsupervised manner. The learned knowledge should be applicable to the test set, even though it is composed of different sets of labels from the training set (see Figure 1.3c).

1.3 Challenges and Solutions

In this section, challenges for the aforementioned problems (Section 1.2.2) are presented, and subsequently corresponding solutions are discussed.



(a) Transductive cross-class transfer learning

(b) Inductive cross-class transfer learning

(c) Unsupervised cross-class transfer learning

Figure 1.3: Three variants of cross-class transfer learning are studied in thesis, and they mainly differ in the learning stage (red dotted area): (a) Transductive cross-class transfer learning – labelled training data and unlabelled test set are accessible, (b) Inductive cross-class transfer learning – only labelled training data is accessible, (c) Unsupervised cross-class transfer learning – only unlabelled training data is available. 'Yes' tick means that datum and corresponding label are given. 'No' tick for no correspondence between datum and label, and it needs to be found in the inference stage.

1.3.1 Transductive Cross-class Transfer Learning

In this variant, it is assumed that the *full* test data is available, but without corresponding labels. The test data can be used to mitigate distribution mismatch between training data and test data.

Challenges Conventionally, a general pipeline to perform cross-class transfer learning is as follows: It is assumed that the training set and the test set are connected in a semantic space. During training process, the first step is to represent seen class labels as a vector in terms of semantic information provided by human expert; the vector is also called a semantic representation or prototype, and it can be considered as a point in the semantic space. Then embedding (a.k.a projection, mapping) function that parametrises the connection between feature space and semantic space is learned with training data and corresponding semantic representation. During testing, unseen class labels are also represented as the semantic representation, and the embedding function is applied without any adaptation to the test samples to project them into the semantic space. Finally, classification is realised in the semantic space (see Section 2.2 for details). In this approach, the learned embedding function is biased towards the training data; thus it may underperform when applied to the test data. In other words, there is a significant *domain gap* between seen classes and unseen classes due to the fact that they are disjoint in terms of label space, making the embedding function less generalisable to the seen classes. This is known as *projection domain shift* problem (Fu et al., 2015a). Hence, the key is how to learn the projection function given labelled training data, and the test set without corresponding labels so that the embedding function is less prone to the domain shift problem (see Figure 1.4).

Solution The main idea is to learn a model that is regularised by semantic representations of unseen classes, belonging to the test set. Note that the correspondences between the test data and their semantic representations are unknown in the training phase. It is assumed that if unseen class semantic representations are incorporated during the model learning, it forces the model not only to respect the structure of the training set, but also the test set. The thesis shows by comprehensive experiments that this helps to alleviate the domain shift problem effectively. The proposed model is built on a dictionary learning/sparse coding, and unsupervised domain adaptation methods: the dictionary learning is mainly used for learning the embedding function in the form of dictionary (Mairal et al., 2009), and unsupervised domain adaptation is used to make the embedding function domain-invariant across the training set and test test (Pan and Yang, 2010; Margolis, 2011).

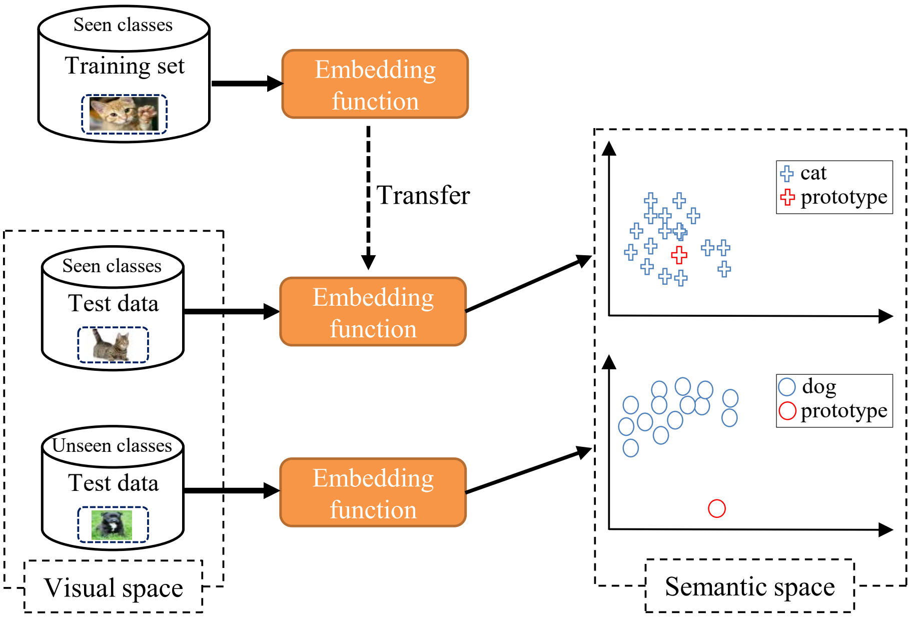


Figure 1.4: An illustration of conventional visual feature embedding approach and how it suffers from the domain shift problem without domain adaptation. When the embedding function (which parametrises the connection between visual space and semantic space) learned using the training set is applied to the test data samples coming from seen classes, it locates them closely to their true semantic representations/prototypes. In contrast, if it is applied to the test data samples coming from unseen classes, then it locates them far away from their true semantic representations/prototypes.

1.3.2 Inductive Cross-class Transfer Learning

Most of the time, in real world scenarios, the test data is only available after the learning stage. This setting considers that only labelled training set is given for the model learning.

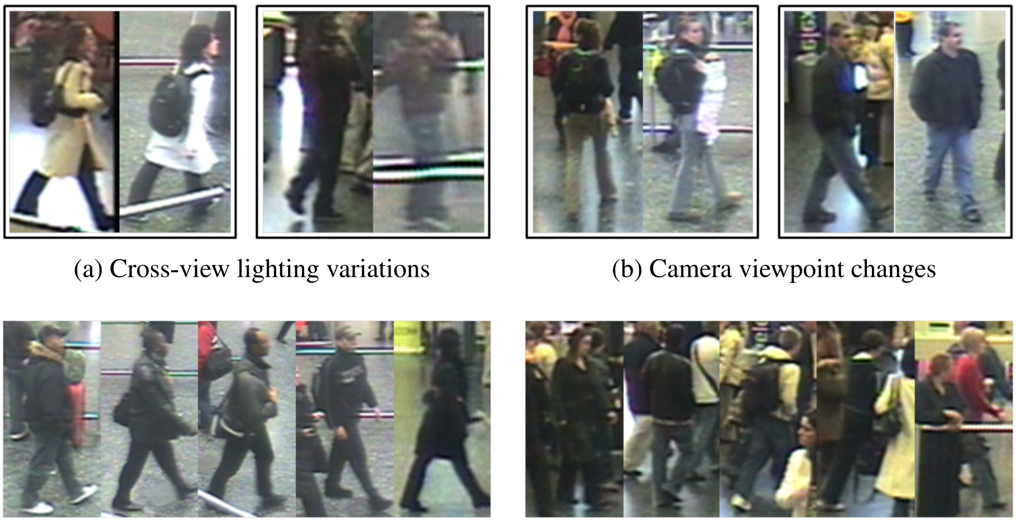
Challenges The main challenge remains the same as the previous variant, that is, the domain shift problem occurs when the model learned only from the training set is applied to the test set. The fact that in this setting the test set is unavailable for the model learning results in even ‘harder’ challenge compared to the previous one. Then, the task is how to utilise the training set *only* so that the learned model is less prone to the domain shift problem.

Solution Motivated by an encoder-decoder paradigm (Bengio et al., 2013), a novel method, named semantic autoencoder (SAE), is proposed. The SAE consists of two functions: an encoder and a decoder. (1) The encoder aims to embed a visual feature vector into the semantic space as in the conventional CCTL models; basically the encoder is the embedding function mentioned above. (2) However, the decoder exerts an additional constraint, that is, the embedding must be able to reconstruct/generate the original visual feature. The additional reconstruction constraint forces the embedding space to preserve richer information; thus the learned embedding function from the seen classes is able to generalise better to the new unseen classes. Moreover, the SAE is designed in a way that the encoder and decoder are linear and symmetric, resulting in an extremely efficient learning algorithm.

1.3.3 Unsupervised Cross-class Transfer Learning

In this variant, an assumption is that no supervision is available for the model learning. In this thesis, unsupervised cross-class transfer learning (UCCTL) refers to the unsupervised setting of person re-identification (ReID) problem, since the setting of UCCTL is the same as that of the person re-identification. Specifically, ReID is a matching problem across non-overlapping cameras: a correct match of a probe image (captured from camera A) needs to be found from gallery images (captured from different cameras such as B, C, D, and E) (see Figure 2.15).

Challenges Performing model learning in an unsupervised manner is intrinsically challenging since any means of additional prior knowledge or supervision is unavailable for helping resolve uncertainty. This is especially true in ReID due to the fact that person images are captured from uncontrollable sources (surveillance cameras) at different locations and time, yielding significant changes in illumination, context, occlusion, background clutters (see examples in Figure 1.5)



(a) Cross-view lighting variations

(b) Camera viewpoint changes

(c) Clothing similarity

(d) Background clutter and occlusions

Figure 1.5: Person re-identification challenges in the wild: (a) Cross-view lighting variations, (b) Camera viewpoint changes, (c) Clothing similarity, (d) Background clutter and occlusions. Two images in each bounding box refer to the same person, but captured in different cameras (Zhu, 2015).

(Zhu, 2015). Therefore, trusting all available visual features extracted from the images blindly for measuring data pairwise similarity is vulnerable to unreliable and noisy features. Conventionally, to eliminate noises from features, most existing supervised methods attempt to learn a subspace which is a lower-dimensional embedding space where the visual similarity relationship is preserved. It is assumed that the subspace preserves discriminative patterns (structure), which are good for differentiating the persons, after projecting the original features into it. However, learning this subspace or discovering discriminative intrinsic structure *without labels* is a key challenge.

Solution Given a unlabelled data the key is to discover some intrinsic structure that is common across different classes so that cross-class transfer learning is possible. This thesis aims to learn common structures in the form of latent attributes so as to facilitate cross-class transfer learning in person re-identification, i.e., the latent attributes are invariant across classes. To this end, robust graph regularised dictionary learning framework is proposed. The graph, which is built from training set, encodes the pairwise connections between different persons across camera views. The dictionary is learned by mining the graph and data; each atom of dictionary represents a particular latent attribute. In the testing stage, the learned dictionary can be used to obtain discriminative representation for a test example.

Remarks Throughout this thesis, transductive cross-class transfer learning and transductive zero-shot classification are used interchangeably. Inductive cross-class transfer learning, cross-class transfer learning, and zero-shot classification are used interchangeably. Similarly, unsupervised cross-class transfer learning refers to the unsupervised setting of person re-identification.

1.4 Contributions

The contributions made in this thesis are summarised as follows:

1. Cross-class transfer learning is formulated as an unsupervised domain adaptation problem. To this end, a regularised dictionary learning-based unsupervised domain adaptation framework is proposed to solve the domain shift problem suffered by existing cross-class transfer learning methods.
2. A novel semantic encoder-decoder model is proposed for inductive cross-class transfer learning. A semantic autoencoder which learns a low-dimensional semantic representation of input data that can be used for data reconstruction is formulated. An efficient learning algorithm is also introduced.
3. A novel graph regularised dictionary learning model is formulated for unsupervised cross-class transfer learning with a new robust l_1-norm graph regularisation term and joint graph and dictionary learning. The method is applied to person re-identification under unsupervised setting (unsupervised cross-class transfer learning). The model only requires unlabelled training data, which makes it suitable for large-scale person re-identification. In addition, an efficient iterative optimisation algorithm is developed for the non-smooth and non-convex objective function of the proposed model. During test time, the model is linear and has a closed-form solution for inference; it is thus very efficient.

1.5 Thesis Outline

This thesis is organised as follows (see also Figure 1.6):

Chapter 2 starts with a brief overview of machine learning tools, and then a review on various existing recognition methods is presented. Also, several common learning strategies which are closely connected to the proposed approaches in this work are discussed.

Chapter 3 explains how transductive cross-class transfer learning can be formulated as unsupervised domain adaptation problem. Specifically, the chapter describes a dictionary learning framework regularised by domain adaptation and visual-semantic similarity constraints. Extensive experiments on four challenging object and action benchmark datasets validate the advantages of the proposed model, and demonstrate that the proposed model outperforms the state-of-the-arts naive transfer learning based models when applied to transductive zero-shot recognition.

Chapter 4 describes inductive cross-class transfer learning approach particularly designed for solving domain shift problem in existing methods in an inductive manner. In particular, a novel semantic autoencoder is proposed based on an encoder-decoder paradigm. Also, efficient optimisation method is presented. Furthermore, extensive experiments are carried out on six benchmarks showing that the proposed SAE model achieves state-of-the-art performance on all the benchmarks.

Chapter 5 presents unsupervised cross-transfer learning approach for person re-identification. The method is based on dictionary learning with robust graph regularisation. In contrast to existing approaches that rely on large number of labelled data, the proposed model does not require any person identity labels, yet it can extract discriminative transferable knowledge from largely potentially noisy visual data. Extensive experiments are conducted on four large benchmark datasets, and the results show that the proposed method significantly outperforms existing unsupervised methods in terms of both matching accuracy and running cost. Furthermore, the proposed model is very flexible in that it can make use of label information if available.

Chapter 6 concludes this thesis and provides a number of directions to be pursued as future work.

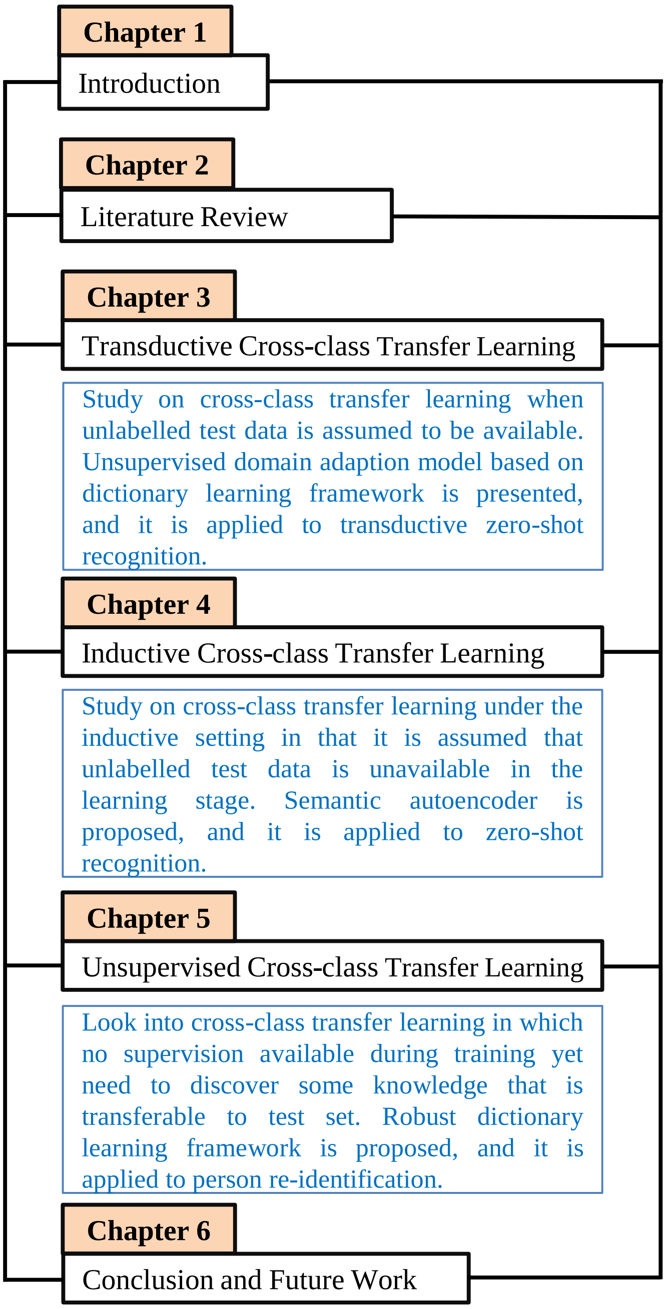

Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Transductive Cross-class Transfer Learning
Study on cross-class transfer learning when unlabelled test data is assumed to be available. Unsupervised domain adaption model based on dictionary learning framework is presented, and it is applied to transductive zero-shot recognition.
Chapter 4
Inductive Cross-class Transfer Learning
Study on cross-class transfer learning under the inductive setting in that it is assumed that unlabelled test data is unavailable in the learning stage. Semantic autoencoder is proposed, and it is applied to zero-shot recognition.
Chapter 5
Unsupervised Cross-class Transfer Learning
Look into cross-class transfer learning in which no supervision available during training yet need to discover some knowledge that is transferable to test set. Robust dictionary learning framework is proposed, and it is applied to person re-identification.
Chapter 6
Conclusion and Future Work

Figure 1.6: A summary of main chapters and structure of all chapters.

Chapter 2

Literature Review

This chapter presents background on several important concepts used throughout this thesis, and review related works. Specifically, Section 2.1 provides background information for understanding general machine learning techniques. Also, transfer learning and visual recognition are reviewed briefly. Then, Section 2.2 provides related works that are closely related to transductive cross-class transfer learning. Similarly, Section 2.3 and Section 2.4 present works related to inductive and unsupervised cross-class transfer learning, respectively. Section 2.2, Section 2.3, and Section 2.4 correspond to Chapter 3, 4, and 5, respectively. Finally, Section 2.5 summarises benchmark datasets used in this thesis.

2.1 Machine Learning Tools and Visual Recognition: An Overview

2.1.1 Machine Learning

The aim of machine learning is to design algorithms that can learn from data. Formally, according to Mitchell et al. (1997), machine learning is ”*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*”. Depending on a specific application, one can design various experiences E, tasks T, and performance measures P. E usually consists of the experience of a set of observed examples encoded in a design matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$, which is also called a data matrix, where N is the number of examples. Each column of the data matrix represents a different example, which is described by d-dimensional feature vector $\mathbf{x}_i$ (i-th column); feature vector is a vector that contains information describing important characteristics of an example. E may also include a corresponding label for each example. The labels for all the examples are represented by a vector, denoted as $\mathbf{y} \in [1, ..., c]^m$, where y_i indicates which of c classes example i belongs to, meaning y_i is associated with a numerical value that corresponds to real-world category. For example, $y_1 = 0$ and $y_2 = 1$: 0 refers to 'dog', and 1 for 'cat'. The label could also become a vector e.g., semantic representation, multi-label (assigning to each sample a set of target labels). The design matrix $\mathbf{X}$, and corresponding labels $\mathbf{y}$ together constitute *a training set*, $\{\mathbf{X}, \mathbf{y}\}$.

As mentioned above, there could be many different tasks T. In this thesis, the following tasks are considered:

- *Classification*: In this task, a goal is to learn a function by using the training set $f : \mathbb{R}^d \rightarrow \{1, ..., c\}$, where the function f takes an example $\mathbf{x}$ ($f(\mathbf{x})$)(e.g., feature vector extracted from an image) as an input, then outputs the category of that example (e.g., 'dog').
- *Verification* : In this task, the algorithm takes the training set and outputs a matching function in which for the given two examples the function is asked to output 'same' or 'different', wherein 'same' indicates that two examples belong to the same classes, 'different' for different classes. For example, in person verification, for given two images, the task is to identify whether they belong to the same class (identity) or different classes (identities).

Each of these tasks has a particular performance measure P. For the classification task, one could define a new set of examples $\mathbf{X}^{(test)}$ and their corresponding labels $\mathbf{y}^{(test)}$, forming a *test set*, and measure classification accuracy of the model based on the testing set. Specifically, first $f(\mathbf{X}^{(test)})$ is performed in that it outputs estimated categories, and then they are compared to the actual $\mathbf{y}^{(test)}$ to find the correct number of matches. Similarly, the test set can be defined for the verification task as well. Then, matching accuracy is calculated using the test set. As a evaluation metric, so called cumulative matching curve (CMC) can be used in which it measures how well the model (a verification system) ranks the identities in the enrolled database with respect to an 'unknown' example.

Generally, depending on the label availability, machine learning approaches are divided into three learning paradigms:

1. *Supervised Learning*: The training set and corresponding labels are available for model learning. This paradigm can be used for the classification and regression tasks. For in-

stance, in the classification the learned model is a function $f(\mathbf{x})$ that maps the examples to class IDs. The proposed approaches in Chapter 3 and Chapter 4 of this thesis fall into this learning paradigm.

2. *Unsupervised Learning*: The training set without corresponding labels is available during model learning. System only relies on unlabelled design matrix. This type of learning is mostly used to discover some hidden structure from the design matrix. In this thesis, this learning paradigm is used to discover visual structure from the unlabelled data in Chapter 5.

3. *Semi-supervised Learning*: The training set with some corresponding labels is available. This paradigm attempts to make use of both worlds the supervised learning and unsupervised learning. In case that the design matrix is *partially* labelled, the system tries to exploit the labelled in addition to unlabelled examples for model learning (Chapelle et al., 2006). This paradigm is not considered in this thesis.

All the learning paradigms are generally treated as an optimisation problem e.g., squared error loss, cross-entropy loss (Bishop, 2006; Goodfellow, 2015). In the supervised learning, for example, the connection between the data examples and their corresponding labels can be parametrised by the objective function. This function is then optimised in the learning stage.

2.1.2 Transfer Learning

The ability of transferring the knowledge learned from training data to test data (which is very different from the training data) is the key factor in transfer learning. In the following (1) a basic idea of transfer learning, and (2) one particular scenario of transfer learning related to this thesis are given.

Transfer Learning vs. Conventional Machine Learning An assumption in conventional machine learning methods discussed in Section 2.1.1 is that the training set and test set are from the same task (i.e., the task is the objective that a model aims to perform such as image classification) and domain (i.e., the domain is where the data is collected from such as images drawn by people) (Figure 2.1). Unfortunately, this assumption fails due to many reasons in real world. For instance, after a model for detecting pedestrians on night-time images is trained, it could be applied to a different domain such as on day-time images. In practice, however, the model's performance deteriorate (often significantly) owing to the fact that the model has biased towards

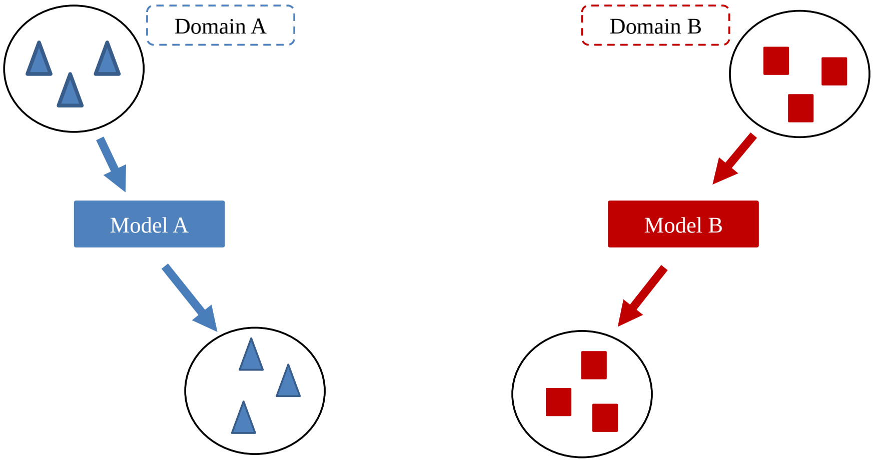


Figure 2.1: A conventional machine learning setting.

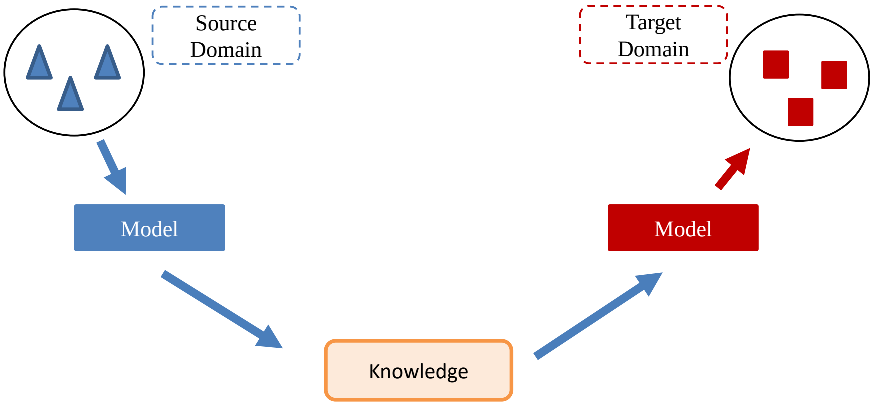


Figure 2.2: An illustration of transfer learning setting.

its training data, thus does not generalise well to the new domain. To deal with such kind of scenario, transfer learning comes to rescue; it attempts to leverage already existing labelled data of related task (Ruder, 2017) i.e., the knowledge is learned from Domain S, and transferred to new domain T (see Figure 2.2).

Formally, given a source domain $\mathcal{D}_S$, and a corresponding source task $\mathcal{T}_S$, and a target domain $\mathcal{D}_T$, and a corresponding target task $\mathcal{T}_T$, the objective of transfer learning is to enable us to learn the target conditional probability distribution $P(\mathbf{Y_T}|\mathbf{X_T})$ in $\mathcal{D}_T$ with the information gained from $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. There are a variety of transfer leaning scenarios depending on how knowledge should be learned, and how the knowledge should be transferred to the target domain. A comprehensive survey on the conventional transfer learning can be found in Pan and Yang (2010). Note that this thesis is different from conventional transfer learning

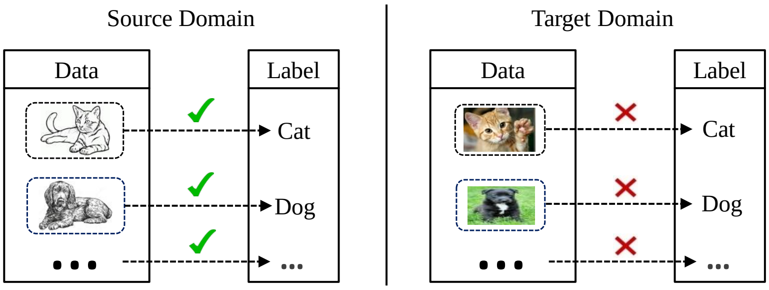


Figure 2.3: An illustration of unsupervised domain adaptation (UDA) setting. In UDA, training set and test set are referred as source domain and target domain, respectively. Similar to conventional supervised learning, label space is the same for source domain and target domain, but the data collected in different (environments) domains. For example, source domain is composed of hand-drawn images, while target domain consists of images captured using high quality cameras.

settings. In all settings mentioned in Pan and Yang (2010), the task is the same for the source domain and target domain; in other words training classes and test classes are joint, not cross-class. To be more specific with regard to the naming of different variants of CCTL, 'transductive' refers to the fact that both training data (labelled) and test data (unlabelled) are considered for model learning, while 'inductive' merely refers to the fact that only labelled training data is used for model learning. Therefore, it is very different from the conventional setting of transfer learning whether it is inductive or transductive (Pan and Yang, 2010).

Unsupervised Domain Adaptation: One particular scenario, which is closely related to the scenario considered in this thesis, is unsupervised domain adaptation (UDA) (Pan and Yang, 2010). In UDA, the source domain and target domain have the same task but different domains i.e., pedestrian detection on night-time and day-time images. The source domain consists of fully labelled data, and target domain is totally unlabelled. The aim of UDA is to learn some knowledge using the labelled source domain, and transfer that knowledge to the target domain. This is accomplished in a transductive manner, meaning both source and target domain data are used for model learning (see Figure 2.3).

A large variety of unsupervised domain adaptation approaches have been proposed ranging from covariate shift, selflabelling, feature representation adaptation, to clustering based approaches (Margolis, 2011). Most of them are designed for text document analysis. However, recently a number of methods are proposed for visual recognition (Ni et al., 2013; Fernando et al., 2013; Patel et al., 2015). Specifically, Ni et al. (2013) proposed to interpolate subspaces through dictionary learning to link the source and target domains. These subspaces are able to capture the intrinsic domain shift and form a shared feature representation for cross-domain recognition. Similarly, Fernando et al. (2013) proposed to learn a mapping function which aligns the source subspace with the target one. This thesis presents an approach that is closely related to this type of transfer learning. However, the presented approach tackles more challenging case. In UDA, the task for source and target domain is same, but only the domains are different. In contrast, in this thesis both tasks and domains are different. Although UDA methods (Ni et al., 2013; Fernando et al., 2013; Patel et al., 2015) are different, they are adapted, and compared with the proposed approach in this thesis (see Table 3.3 in Chapter 3).

2.1.3 Visual Recognition

As mentioned in Section 1.1, the visual recognition can be divided into two types: category recognition and instance recognition. Depending on a particular task the category recognition could take different forms (comprehensive review of different visual recognition tasks is given in Grauman and Leibe (2011)). Specifically, two category recognition tasks considered in this thesis are as follows:

- *Object*[1] *recognition* is a process for identifying a category of an object in visual data e.g., an image or video.
- *Action recognition* is a process of recognising a particular action in video. The action could be ‘jumping’, ‘riding’, and ‘swimming’.

Similarly, for instance recognition, this thesis deals with instance(-level) verification problem.

Conventional Algorithms for Visual Recognition: In both types, most contemporary recognition methods follow a common pipeline, which is depicted in Figure 2.4. Overall, it consists of two main steps: (1) feature extraction, and (2) classification/categorisation model learning. Both steps are detailed below:

1. *Feature Extraction*: Visual feature vector (a.k.a feature representation) is a vector that contains information describing important characteristics of an object in relation to other objects. Due to the fact that images are a very high dimensional, identifying what features discriminative and useful is challenging. In the literature there are many kinds of

[1] Note that ‘object’ refers to ‘person’, ’scene’, ‘animal’, or a any material thing that can be seen and touched.

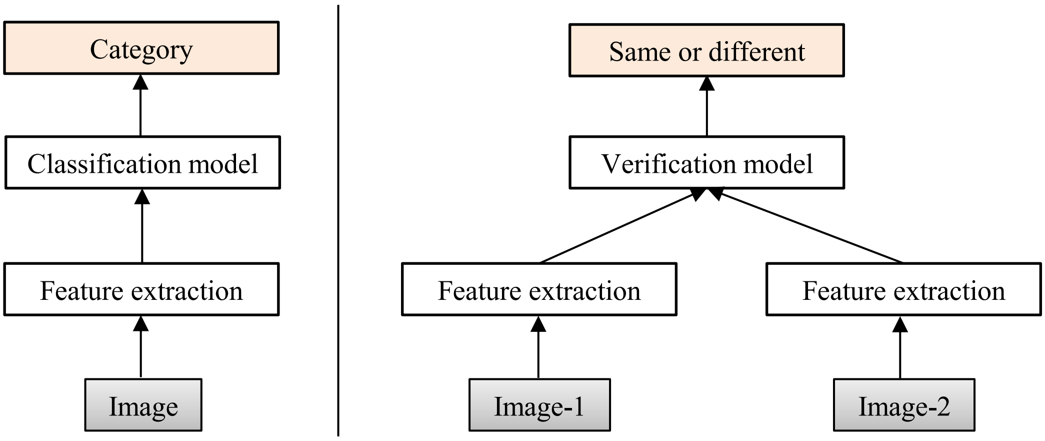


Figure 2.4: Pipeline for category recognition/classification (left), and instance verification (right). Category recognition is a classification problem: for a given input image, the system should output the category of that image. In instance verification, for given two images (Image-1 is a probe image and Image-2 is from gallery), a system should identify whether two images belong to the same class or not.

approaches were proposed. Overall, they can be divided into two groups: (1) 'window-based' representation and (2) 'part-based' representation (Grauman and Leibe, 2011). The window-based approaches take a whole image or a region of interest ('window') from the image, and output a single feature vector (descriptor), summarising appearance of the image in terms of texture, shape, color and geometric cues. Examples include Histogram of Gradients (HOG) (Dalal and Triggs, 2005) (see Figure 2.5), Scale-invariant Feature Transform (SIFT) (Lowe, 1999), Local Binary Pattern (LBP) (Ojala et al., 2002) and many variants of these methods. For second group, they first define parts of the particular object (e.g., human can be decomposed into different parts such as head, shoulders, hand, ...), and then for each part separate descriptor is obtained, and finally all descriptors are combined. Examples include Star Model (Szummer and Picard, 1996), Pictorial Structure Model (Felzenszwalb and Huttenlocher, 2005), and Sparse Flexible Model (Carneiro and Lowe, 2006). The extracted features in both types are known as 'hand-crafted' (hand-engineered or low-level) features, since these approaches require a great deal of supervision from human during the feature extraction. Both groups have their strength and weaknesses. For example, in terms of time complexity, window-based approaches are faster. In terms of robustness against occlusion and some background 'clutter', part-based approaches are better. Main weakness of part-based approaches, however, is that defining parts is problematic in real-world images Grauman and Leibe (2011). Formally, at this stage, f_1 function is designed so that feature vector $\mathbf{x}$ can be obtained for a given image $\mathbf{I}$, $f_1 : \mathbf{I} \rightarrow \mathbf{x}$.

2. *Model Learning*: (1) Category recognition (classification): After having the feature extraction function, it applies to all images in the training set, and test set. Using the training set with extracted features, one can learn a classification model by taking off-the shelf classification methods such as Support Vector Machines (Cortes and Vapnik, 1995) or Nearest Neighbour classifier. Finally, the learned model is applied to test examples to categorise them. Formally, at this stage the goal is to learn a function f_2 that classifies the feature vector $\mathbf{x}$ extracted from image $\mathbf{I}$ into a particular category $f_2 \ : \ \mathbf{x} \ \rightarrow \ c$, where c indicates a particular category e.g., 'cat'. (2) Instance verification: Similar to the classification, features are extracted for training set. Then, distance metric learning approaches are applied to learn distance metric using the training set. The key idea of the metric learning is to learn the similarity function from the data using some closeness constraints. Ideally, the metric should bring examples that have the same category, while pushing far away that of different categories. There are various approaches ranging from informaintro: benchmarkstion-theoretic metric learning (Davis et al., 2007), large margin nearest neighbour classifier (Domeniconi et al., 2005), to regressive virtual metric learning (Perrot and Habrard, 2015). Formally, the goal is to learn a function f_2 that identifies whether given two images' feature vectors $(\mathbf{x}_1, \mathbf{x}_2)$ belong to the same category or different categories, essentially a matching problem $f_2 \ : \ (\mathbf{x}, \mathbf{x}) \ \rightarrow \ 'Same' \ or \ 'Different'$, where $'Same'$ indicates they belong to the same class, $'Different'$ otherwise.

Deep Learning for Visual Recognition: Recently, deep learning approaches become significantly successful for the visual recognition. In terms of performance it is much better than the 'shallow' methods Krizhevsky et al. (2012). Deep learning is a class of approaches that employ artificial neural networks (ANN) with multiple layers of increasingly richer functionality. This is also known as deep neural networks (DNN). The name 'deep' is owing to the fact that the multiple layers are used in the network (see Figure 2.7). One of the successful types of DNN in computer vision research is deep convolutional neural network (CNN). There are a wide variety of CNNs were proposed including AlexNet (Krizhevsky et al., 2012), VGG (see Figure 2.7) (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), Inception (Szegedy et al., 2017), and ResNet (He et al., 2016). Importantly, deep learning methods learn feature extraction and clas-

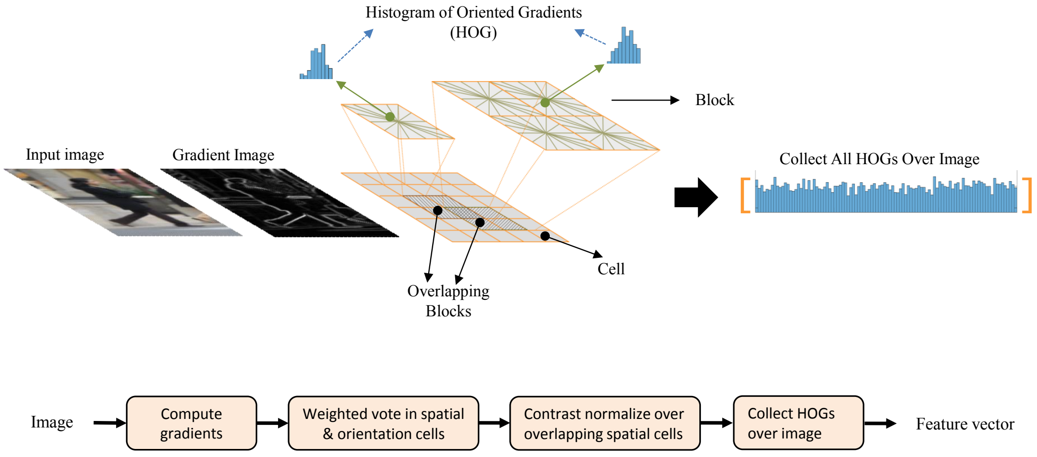

Histogram of Oriented Gradients (HOG)
Block
Input image
Gradient Image
Collect All HOGs Over Image
Cell
Overlapping Blocks
Image
Compute gradients
Weighted vote in spatial & orientation cells
Contrast normalize over overlapping spatial cells
Collect HOGs over image
Feature vector

Figure 2.5: Pipeline for HOG feature extraction (Dalal and Triggs, 2005).

sification model in an end-to-end manner, unlike the conventional '2-step' methods (see Figure 2.6). Formally, in CNN, f function is learned so that f takes the image $\mathbf{I}$ as input, and echoes the category of the image $\mathbf{I}$, $f : \mathbf{I} \rightarrow c$, where c indicates a particular category e.g., 'cat'. Similarly, it can be applied to the verification. Representative approaches include Siamese network (Koch, 2015), Triplet network (Hoffer and Ailon, 2015).

One of the advantages of CNN is that it can be used as a generic feature extraction function. That is, first CNN is trained with very large dataset consisting of millions of observing examples in a supervised manner. During the training, CNN learns many common patterns across many categories. Then, the trained model is used as a generic feature extraction function to extract features for an image. The features extracted by CNN is often referred to as 'deep' features.

This thesis does not focus on feature extraction component, rather largely focus on model learning for classification/verification. Therefore, both hand-crafted (window-based representation) and deep features used in this work are extracted by the methods mentioned above.

2.2 Transductive Cross-class Transfer Learning

The goal of the cross-class transfer learning for category recognition is to learn some knowledge from the training set that can be applied to categorise data from the test set, which has different labels from the training set. A common pipeline for performing CCTL for category recognition is to establish a joint embedding (semantic space) for visual features and labels, and then per-

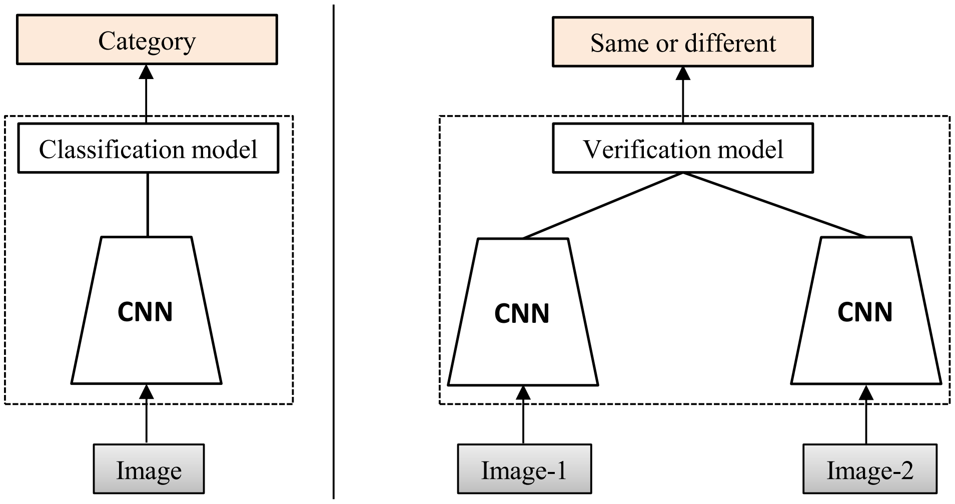


Figure 2.6: Pipeline for category recognition/classification (left), and instance verification (right) using deep neural networks.

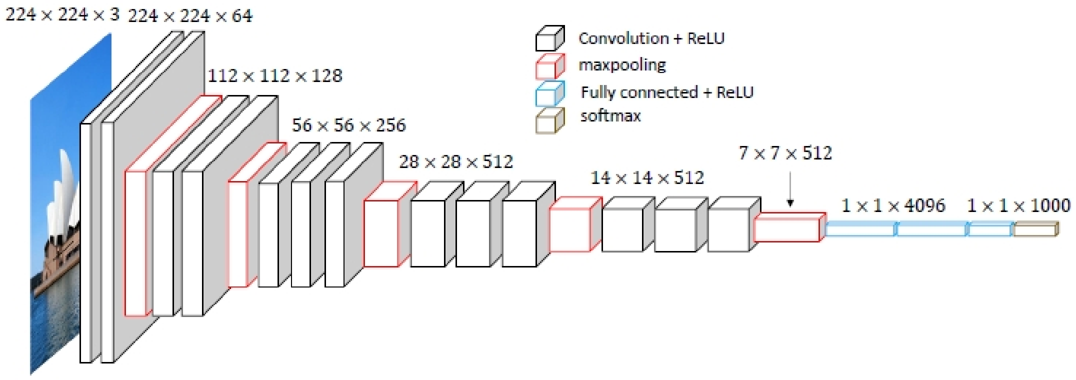


Figure 2.7: An example of a deep learning model: VGG network (Simonyan and Zisserman, 2014). Any deep model architecture consists of several combinations of convolution, ReLU, maxpooling, fully connected, and softmax layers.

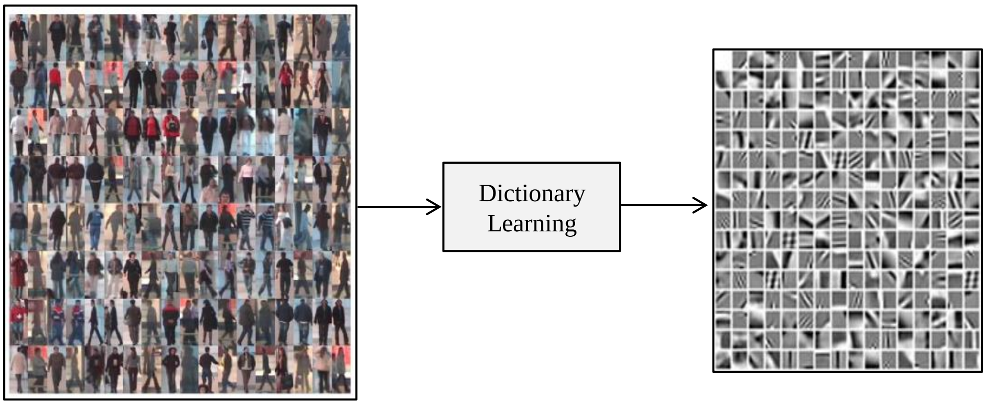


Figure 2.8: An illustration of dictionary learning: given training data, it learns hidden patterns/structures from the data in an unsupervised manner.

forming nearest neighbour search Lazaridou et al. (2014). In the following, first an overview for dictionary learning for sparse coding method (1) is presented, because Chapter 3 presents a method that is build on this technique. Then, in order to present whole pipeline clearly for performing CCTL for category recognition, the concept of label embedding (2) followed by a linear regression-based cross-class transfer learning method (3) is presented; the projection domain shift problem (4) is also discussed. Finally, contemporary CCTL methods for category recognition are discussed (5).

1) Dictionary learning for sparse coding: an overview

Dictionary learning is a method whose aim is to learn a set of basis vectors that can be used to best approximate the data; the learned basis vectors are referred to as a dictionary. It is mainly used to discover important hidden structure from the data. In fact, widely used k-means clustering can be considered as a naive dictionary learning method, in which cluster centroids correspond to the atoms of the dictionary. Formally, the dictionary learning is formulated (with proper matrix dimensions) as follows:

$$\min_{\mathbf{D},\mathbf{Y}} ||\mathbf{X}-\mathbf{D}\mathbf{Y}||_F^2 + \alpha||\mathbf{Y}||_1 \tag{2.1}$$

where $||\mathbf{X}-\mathbf{D}\mathbf{Y}||_F^2$ is the reconstruction error term evaluating how well a linear combination of the learned atoms of dictionary $\mathbf{D}$ can approximate the input design matrix $\mathbf{X}$, and $||\cdot||_F$ denotes the matrix Frobenious norm; $||\cdot||_1$ is the sparsity term favouring small number of atoms to be used for reconstruction; this term is weighted by α, and $\mathbf{Y}$ is a sparse representation. The dictionary learning is closely connected to a sparse coding (Wright et al., 2010). In the sparse coding, a

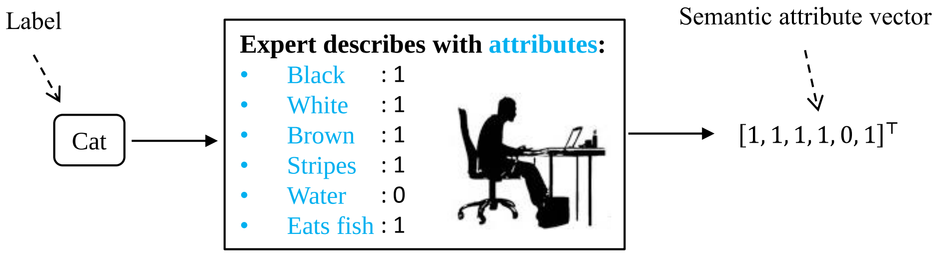


Figure 2.9: Label embedding: attributes are used. Basically, the label is converted to a vector representation based on a pre-defined set of attributes defined by an expert.

goal is to obtain sparse representation based on pre-defined dictionary. In contrast, in dictionary learning, sparse representation and the dictionary are learned simultaneously. Both terms are used interchangeably in this thesis.

The main advantage of dictionary learning framework is that it is easy to adapt to a new problem at hand. Therefore, there are a variety of methods ranging from unsupervised, semi-supervised dictionary learning to supervised dictionary learning methods. Guo et al. (2012) proposed a method to learn discriminative dictionary learning for face verification with pairwise constraints. Patel et al. (2014) applied the dictionary learning to face recognition. Wang et al. (2016b) integrated the dictionary learning and deep learning. Beyond recognition tasks, it was applied to tracking (Yang et al., 2014a), super resolution (Yang et al., 2012), compression (Nejati et al., 2016) and more.

2) Semantic label embedding

Label embedding is used in all existing CCTL methods for category recognition. At higher level, the label embedding is where human knowledge is represented. It is crucial for any existing CCTL methods for category recognition because this helps to connect between seen classes and unseen classes. In simple terms, the label embedding is transforming label name into a detailed description based on a predefined ontology. Say ’cat’ is a label of a particular image, and it can be described with a set of attributes such as ’is black’, ’is white’, ’has strips’, ’eats fish’ and the others. When attributes are used, this type of label embedding is referred to as *attribute label embedding* (Lampert et al., 2009). More formally, ’cat’ label name is transformed to a binary vector representation that consists of an array of numbers consisting of $\{0,1\}$, $'1'$ indicates the presence of particular attribute, $'0'$ otherwise i.e., ‘cat’ → $[0,0,0,...,1,1]$. The number of attributes, ontology, is determined by an expert as shown in Figure 2.9. In literature, this vector is also called a semantic vector, because it captures semantically meaningful attributes. Also, the semantic vector is considered as a point in the semantic space, and it is called a *prototype* (or a *signature* (Romera-Paredes and Torr, 2015)). This is illustrated in Figure 2.10. Importantly, in the label space, there is no notion of distance between labels, whereas the distance between semantic vectors can be calculated in the semantic space. For example, it can be clearly seen from Figure 2.10 that distance between 'cat' and 'dog' is less than that of between 'cat' and 'horse' indicating 'cat' is more similar to 'dog' than 'horse' in terms of attributes.

Candidate semantic spaces: The semantic embedding spaces considered by most early works are attribute spaces (Lampert et al., 2009; Liu et al., 2011b; Mensink et al., 2014; Jayaraman and Grauman, 2014; Akata et al., 2013; Wang and Ji, 2013). However, as mentioned, to represent an object class in an attribute semantic space, an attribute ontology has to be defined manually (e.g., what attributes are needed to describe different types of animals) and each class needs to be annotated by an attribute vector (e.g., an expert needs to define various attributes as shown above). Such requirements hinder the scalability of an attribute semantic space based CCTL method. To overcome this, more recent works explore the semantic word vector space (Frome et al., 2013; Fu et al., 2015c), which is learned using large corpus of unannotated text for natural language processing tasks such as sentence completion (Mikolov et al., 2013). The text corpus is so big that any class label or textual description of the class can be embedded in this space, effectively mitigating the scalability issue. Beyond semantic attribute or word vector, direintro: benchmarksct learning from textual descriptions of categories has also been attempted, e.g., Wikipedia articles (Elhoseiny et al., 2013; Lei Ba et al., 2015), sentence descriptions (Reed et al., 2016a). The semantic word and attribute label embeddings are exploited in this thesis.

Equipped with the understanding of the label embedding, the semantic vector, and the semantic space, regression based cross-class transfer learning method is presented in the following.

3) Cross-class transfer learning by linear regression

Cross-class transfer learning can be realised by two steps (Lazaridou et al., 2014):

1) Learning stage: an embedding function from a visual feature space to a semantic space is learned using the training set. The semantic space plays a role of bridge (as a knowledge transfer) that connects the training set (seen) classes and the test set (unseen) classes.

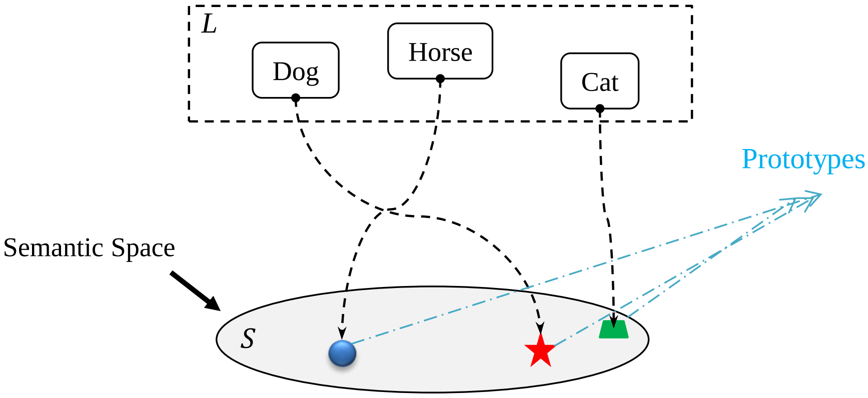


Figure 2.10: A semantic space. The labels named 'Dog', 'Horse', and 'Cat' are embedded to the semantic space S. The embedding of each class label constitutes a point in the semantic space, and they are called prototypes.

2) Testing stage: the learned embedding function is transferred to the test set so that test data can be embedded into the semantic space using this function, and then classification is realised by simply performing nearest neighbour classification in the semantic space.

Formally, assume a labelled training set is given as $\mathcal{D}_{tr} = \{\mathbf{X}_{tr}, \mathbf{y}_{tr}, \mathbf{S}_{tr}\}$, where $\mathbf{X}_{tr}$ is a design matrix, $\mathbf{y}_{tr} = \{y_i\}_{i=1}^{m}$ and $\mathbf{S}_{tr}$ are corresponding labels and a semantic representation matrix, respectively. Similarly, $\mathcal{D}_{te} = \{\mathbf{X}_{te}, \mathbf{y}_{te}, \mathbf{S}_{te}\}$ is for the test set. In both sets each class label y_i is associated with a pre-defined semantic vector $\mathbf{s}_i$, referred to as prototype. For clarity, it is assumed that all matrix and vectors have a proper dimension, thus they are omitted. In the learning stage, using the training set, an embedding function embedding the example features to semantic space is learned. During deployment, $\mathbf{y}_{te}$ has to be estimated for $\mathbf{X}_{te}$. Note that the training classes (seen) and test classes (unseen) are disjoint: $\mathbf{y}_{tr} \cap \mathbf{y}_{te} = \varnothing$. Learning the embedding function can be formulated as an optimisation problem by a linear ridge-regression as follows:

$$\min_{\mathbf{W}} \|\mathbf{X}_{tr}\mathbf{W} - \mathbf{S}_{tr}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \tag{2.2}$$

$\mathbf{W}$ is the embedding function, and λ is a regularisation parameter. This formulation has a closed-form solution. After learning the embedding function $\mathbf{W}$ using Eq. (2.2), a new test example can be embedded $\mathbf{x}_j$ into the semantic space as follows:

$$\hat{\mathbf{s}}_j = \mathbf{x}_j \mathbf{W} \tag{2.3}$$

After that, the classification/categorisation of the test data in the semantic space can be achieved

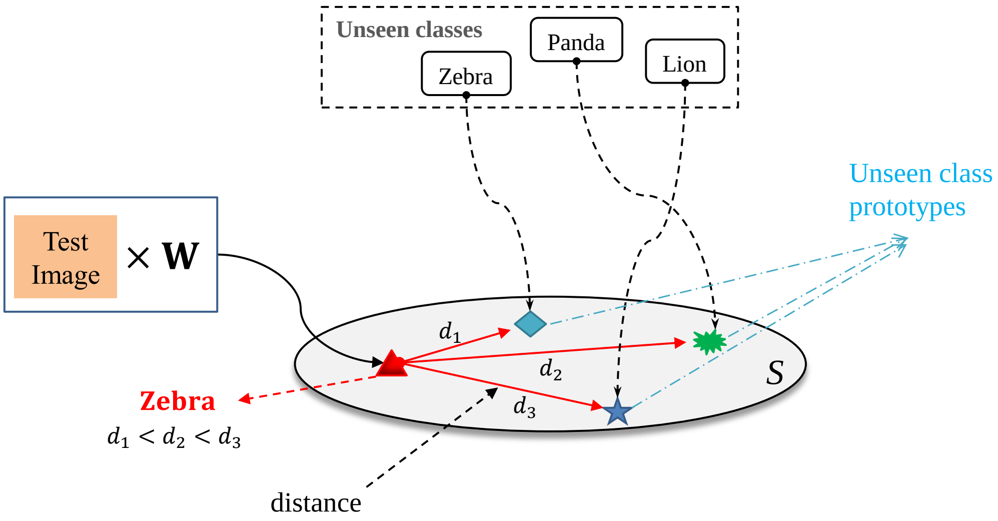


Figure 2.11: Cross-class transfer learning for category recognition.

by simply calculating the distance between the estimated semantic representation $\hat{\mathbf{s}}_j$ and test class prototypes $\mathbf{S}_{te}$:

$$\Phi(\mathbf{x}_j) = \arg\min_k D(\hat{\mathbf{s}}_j,\ \mathbf{s}_k) \tag{2.4}$$

where D is a distance function, $\mathbf{s}_k$ is the semantic representation of the k-th test class label, and $\Phi(\cdot)$ returns the class label of a sample. This procedure is illustrated in Figure 2.11.

4) Projection domain shift

Inherently, existing CCTL methods including regression-based method presented above suffer from projection domains shift problem (Fu et al., 2014) in which the learned function is biased towards the training set, thus it may underperform for the test data. Figure 2.12 depicts conventional visual feature embedding approach and how it suffers from the domain shift problem. For the two classes in the source domain (training data) and the two in the target (testing data), both their visual feature vectors and class names are embedded in a semantic attribute space shared between the two domains. When the feature embedding function is learned from the source and applied without adaptation to the target, the target domain data and their class prototypes are well separated, resulting in poor classification. This is due to domain shift – although both tiger and zebra have the 'has stripe' attribute, their stripes are visually very different. To alleviate this problem, recently transductive cross-class transfer learning methods, which are discussed in the next section, are proposed. In this thesis, Chapter 3 and Chapter 4 are devoted to overcome the

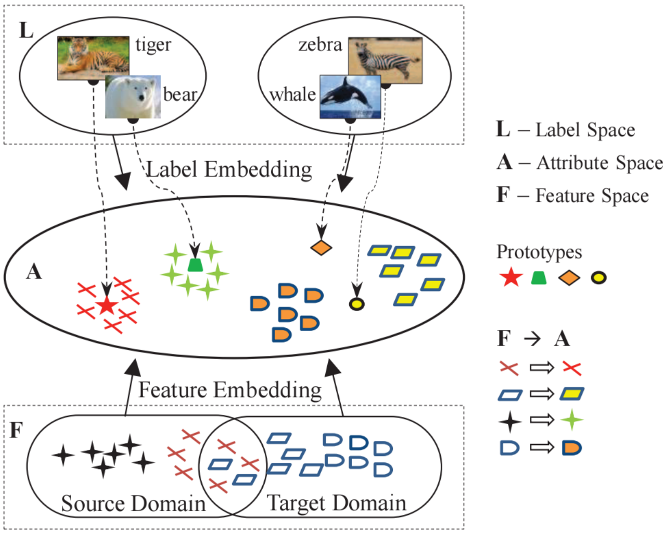


Figure 2.12: An illustration of conventional visual feature embedding approach and how it suffers from the domain shift problem without domain adaptation. For the two classes in the source domain and the two in the target, both their visual feature vectors and class names are embedded in a semantic space (attribute in this case) shared between the two domains. When the feature embedding function is learned from the source and applied without adaptation to the target, the target domain data and their class prototypes are well separated, resulting in poor classification. This is due to domain shift – although both tiger and zebra have the 'has stripe' attribute, their stripes are visually very different.

domain shift problem.

5) Transductive cross-class transfer learning

There are a few transductive cross-class transfer learning approaches in the literature. Fu et al. (2014) proposed a heuristic one-step self-training strategy to pull the prototype towards its closest data points (not necessarily from the same class) followed by a multi-view embedding based on canonical correlation analysis to align different semantic spaces with the feature space. Similarly, Guo et al. (2016a) proposed shared model space to align source data and target data. All these methods indeed help to resolve the domain shift problem to some extent. Both approaches are fall into discriminative models. This thesis also proposes an approach to pursue this direction. Specifically, a novel regularised dictionary learning framework inspired by the unsupervised domain adaptation methods is introduced. Also, for the first time, the dictionary learning framework is uniquely applied to cross-class transfer learning, unlike Fu et al. (2014) and Guo et al. (2016a) that use canonical component analysis, and label embedding approach, respectively.

2.3 Inductive Cross-class Transfer Learning

The plan for this section is as follows: 1) first linear autoencoder is reviewed, because Chapter 4 presents a method that is based on this. 2) Then, related methods/concepts to inductive cross-class transfer learning are given.

1) Autoencoder: an overview

There are many variants of autoencoders in the literature. They can be roughly divided into two groups which are (1) undercomplete autoencoders and (2) overcomplete autoencoders. In general, undercomplete autoencoders are used to learn the underlying structure of data and used for visualisation/clustering. In contrast, overcomplete autoencoders are used for classification based on the assumption that higher dimensional features are better for classification.

In its simplest form, an autoencoder is linear and only has one hidden layer shared by an encoder and a decoder. The encoder projects the input data into the hidden layer with a lower dimension (i.e., higher dimension for the overcomplete autoencoders) and the decoder projects it back to the original feature space and aims to faithfully reconstruct the input data. Formally, given an input design matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$, it is projected into $k-$dimensional latent space with a projection matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$, resulting in a latent representation $\mathbf{WX} = \mathbf{h} \in \mathbb{R}^{k \times N}$. The obtained latent representation is then projected back to the feature space with a projection matrix $\mathbf{W}^* \in \mathbb{R}^{d \times k}$ and becomes $\hat{\mathbf{X}} \in \mathbb{R}^{d \times N}$, having $k < d$. i.e., the latent representation reduces the dimensionality of the original data input. The aim is to minimise the reconstruction error, i.e., $\hat{\mathbf{X}}$ is as similar as possible to $\mathbf{X}$ (see Figure 2.13). This is achieved by optimising the following objective function:

$$\min_{\mathbf{W},\mathbf{W}^*} ||\mathbf{X} - \mathbf{W}^*\mathbf{WX}||_F^2 \tag{2.5}$$

From this formulation, one can design many different variants. For example, non-linearity can be introduced by $\phi(\mathbf{WX})$, where ϕ is a non-linear function, or regularisation can be added to the objective $\Omega(\mathbf{WX})$, where Ω could be a sparsity constraint (Goodfellow et al., 2016). In this thesis, the standard linear autoencoder is extended by introducing a novel regularisation so that cross-class transfer learning is possible (see Chapter 4).

Autoencoder is only one realisation of the encoder-decoder paradigm. Recently deep encoder-decoder has become popular for a variety of vision problems ranging from image segmentation

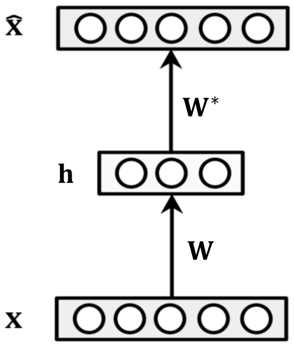


Figure 2.13: A linear autoencoder: $\mathbf{X}$ is a design matrix, $\mathbf{h}$ is a hidden space, and $\hat{\mathbf{X}}$ is reconstruction. $\mathbf{W}$ is an encoder function, while $\mathbf{W}^*$ for a decoder function.

(Badrinarayanan et al., 2015) to image synthesis (Yan et al., 2016; Reed et al., 2016b). Among them, a few recent works proposed conditional autoencoder, which is closely related to the approach proposed in Chapter 2 by introducing label information into the latent embedding space shared between the encoder and decoder (Yan et al., 2016; Reed et al., 2016b). There are several differences compared to (Yan et al., 2016; Reed et al., 2016b): (1) In terms of goal/task: Their goal is directed to more of a generative task. For example, (Yan et al., 2016) try to generate an image given a semantic attribute vectorm while Reed et al. (2016b) try to generate image using text for the same purpose. In contrast, this thesis focuses on a discriminative task, namely classification and verification tasks; (2) In terms of model architecture and optimisation: the methods proposed in this thesis are not based on deep learning, therefore the objective functions and the proposed optimisations for them are very different. Certainly, these methods can be applied to the methods proposed in this thesis to generate samples for the training purposes so that the model can generalise more by taking advantage of synthesised images. These are left for future work.

2) Inductive cross-class transfer learning

The assumption of test data availability is often invalid in the context of CCTL because new classes typically appear dynamically and unavailable before model learning. Therefore, most existing methods focus on the inductive setting of CCTL in which test data is unavailable during a model learning. Existing CCTL models differ in how embedding from the visual space to the semantic space (embedding function) is established. They can be divided into three groups as depicted in Figure 2.14. Specifically, (1) Methods in the first group learn the embedding function from a visual feature space to a semantic space either using conventional regression (as shown

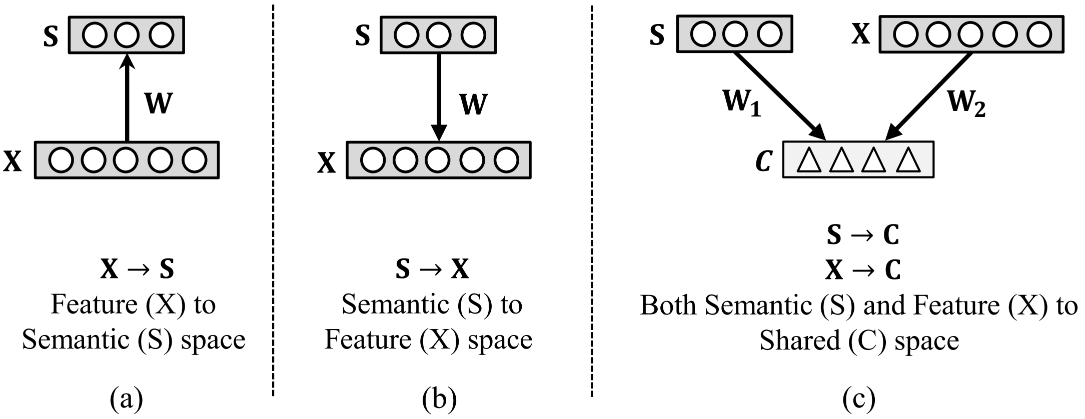


Figure 2.14: Different ways of embedding function learning.

above) or ranking models (Lampert et al., 2009; Akata et al., 2015) or via deep neural network regression or ranking (Socher et al., 2013; Frome et al., 2013; Reed et al., 2016a; Lei Ba et al., 2015). (2) The second group chooses the reverse projection direction, i.e., learning embedding function from semantic space to the visual space (RRZSL – RR stands for reverse regression). The motivation is to alleviate the hubness problem that commonly suffered by nearest neighbour search in a high dimensional space (Shigeto et al., 2015). The hubness problem refers to the fact that learned model frequently predicts the same labels ('hubs'). By changing direction, the label distribution with smaller variance can be achieved helping allivate the hubness problem. (3) The third group of methods learn a common space where both the feature space and the semantic space are projected to (Lu, 2015; Zhang and Saligrama, 2016; Chao et al., 2016a). The proposed method in this thesis (see Chapter 4) is built on an encoder-decoder paradigm in which it consists of an encoder and a decoder function. The encoder in the model is analogous to the first group of models, whilst the decoder does the same job as the second group. The proposed method can thus be considered as a combination of the two groups of models.

2.4 Unsupervised Cross-class Transfer Learning

This section discusses unsupervised cross-class transfer learning (UCCTL) methods in which labels are unavailable during model learning. This type of learning is closely related to clustering. Often, the clustering is used to discover some intrinsic common structure from the data. Similarly, unsupervised CCTL methods are designed to discover discriminative transferable features from the unlabelled data. Under this setting classification or regression cannot be performed, since any kind of label information or ontology during the training is unavailable to connect between training set classes and test set classes unlike two previous variants of CCTL. However, this setting can be applied to instance verification, since verification is a matching problem that cares largely similarity metric function, thus class information/ontology is not required.

The structure of this section is as follows: (1) first conventional graph regularisation is presented, because Chapter 5 presents a method that is build on this regularisation. (2) Then, person re-identification, which is considered as UCCTL problem in this thesis as well as related work are discussed.

1) Graph regularisation: an overview

A graph regularisation is one of the most common regularisation terms in data representation learning frameworks such as dictionary learning, autoencoder, and (robust) principal component analysis. When this regularisation is integrated into model learning, it enforces the model to respect the local data geometric structure (manifold). An idea of the graph regularisation is from the spectral graph theory and manifold learning theory (Chung, 1997; Belkin and Niyogi, 2003). Formally, let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be a sparsely connected undirected graph (a.k.a a nearest neighbour graph) between a set of data points where $\mathbf{V}$ is a set of graph vertices representing the data points and $\mathbf{E}$ the edge set. This graph can be encoded by a weight matrix $\mathbf{Q} \in \mathbb{R}^{N \times N}$ for N data points in which the matrix $\mathbf{Q}$ characterises the geometric structure of the data. The weight $\mathbf{Q}_{ij}$ can be defined in several ways:

1) *0-1 weighting*: $\mathbf{Q}_{ij} = 1$ if nodes i and j are connected by an edge.

2) *Heat kernel weighting*: If nodes i and j are connected,

$$\mathbf{Q}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}} \tag{2.6}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are data points.

3) *Dot-product weighting*: If nodes i and j are connected,

$$\mathbf{Q}_{ij} = \mathbf{x}_i^\top \mathbf{x}_j \tag{2.7}$$

Mathematically, the graph regularisation term $\Omega(\mathbf{V})$ is defined as:

$$\Omega(\mathbf{V}) = \sum_{ij}^{N} \mathbf{Q}_{ij} \|\mathbf{v_i} - \mathbf{v_j}\|_2^2 \tag{2.8}$$

where $\|\mathbf{v_i} - \mathbf{v_j}\|_2^2$ measures distance between samples, $\mathbf{v}_i$ is i-th column of $\mathbf{V}$ that needs to be learned. Intuitively, Eq. (2.8) means that data points $\mathbf{V}$ in a some space should be smooth with regards to the graph, that is, their distances need to conform to the visual similarity relationship embedded in the graph i.e., $\mathbf{Q}$ is usually constructed using hand-crafted features before model learning, while $\mathbf{V}$ needs to be learned with respect to $\mathbf{Q}$.

Recently it has been shown that regardless what unsupervised learning model is taken, a method can benefit significantly from introducing the graph regularisation term in its objective function. For instance, Zheng et al. (2011) proposed to include the graph regularisation into the dictionary learning. Similarly, Liao et al. (2017) considered to include this term into the autoencoder objective, and Jiang et al. (2013) integrated this regularisation into principal component analysis (Jiang et al., 2013). Moreover, it is extensively used for semi-supervised learning and label propagation (Zhu, 2005). For instance, in semi-supervised learning, when partial labelled data are available, the graph can be constructed in a supervised manner with the labelled data, while the remaining data are used to construct the graph in a unsupervised manner. This way, during model learning the knowledge may be propagated from the labelled data to unlabelled data resulting in a better model.

Despite the success of the graph regularisation term, this thesis found two critical limitations of it. The first one is the use of squared l_2 distance which is prone to outliers and noise. The next one is use of the weight matrix $\mathbf{Q}$ for the model learning, that is, existing methods assume that $\mathbf{Q}$ captures true manifold of the data. However, the fact that $\mathbf{Q}$ is constructed from noisy features that are contaminated by noise and outliers results in noisy weight matrix that may not capture true manifold structure of the data well. Solutions to both of them are presented in this thesis (see Chapter 5).

2) Person re-identification (ReID)

Person re-identification (ReID) is a fundamental problem in surveillance system, and its aim is to match people across non-overlapping camera views distributed at different physical locations (Gong et al., 2014) (see Figure 2.15). Depending on the availability of images across camera views, ReID can be performed in a singe-shot Xiong et al. (2014b); Wei et al. (2017); Xiao

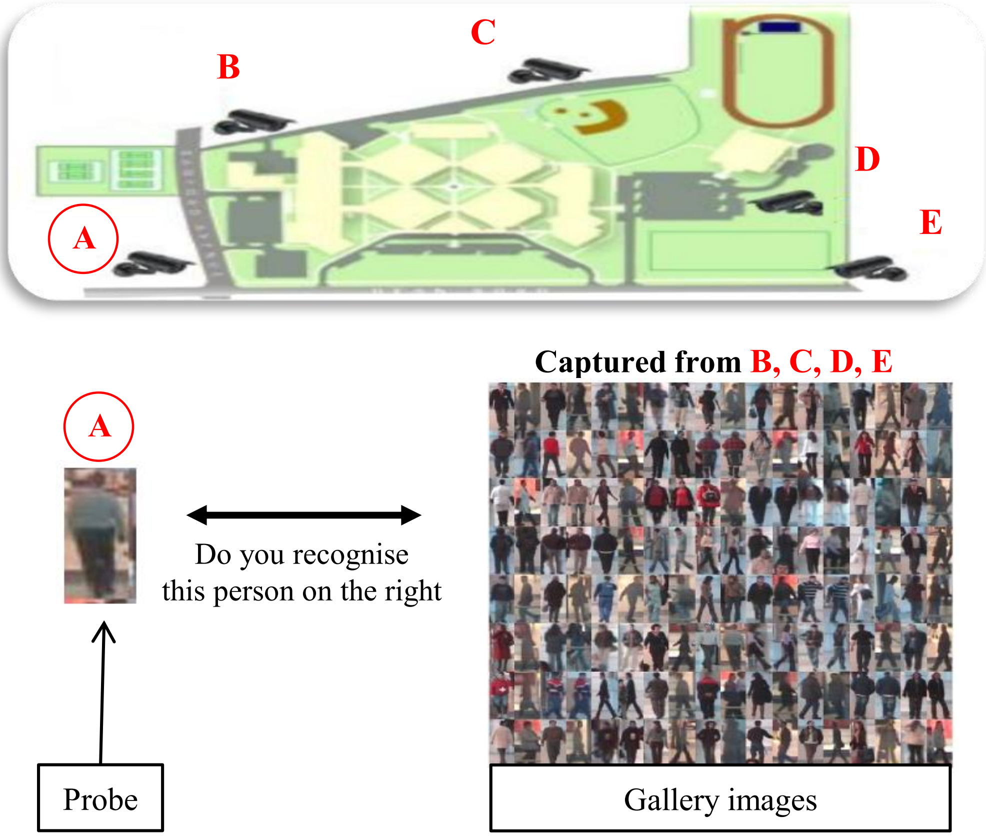


Figure 2.15: Person re-identification: a correct match of a probe image (from A) needs to be found from gallery images from B, C, D, and E. A, B, C, D, and E are non-overlapping cameras.

et al. (2016) (i.e., only one image per camera view) or a multi-shot manner Ma et al. (2017); McLaughlin et al. (2016). In the following, first a number of conventional ReID studies under unsupervised setting are discussed. Then, contemporary unsupervised deep learning methods for ReID are presented. Finally, supervised ReID methods are presented briefly. More thorough reviews can be found in (Zheng et al., 2016; Gong et al., 2014; Vezzani et al., 2013; Bedagkar-Gala and Shah, 2014).

Conventional unsupervised learning for ReID. Generally ReID system consists of two key components, like any visual conventional recognition system: (1) feature representations of persons, and (2) a matching model. These components are detailed below.

1. Feature Representation In ReID, designing a discriminative feature representation that is robust against large cross-view changes, illumination, background clutter and occlusion is challenging. Various feature representations have been designed based on colour, texture, gradient, edge, pose, and shape cues (Gray and Tao, 2008; Farenzena et al., 2010; Hirzer et al., 2012; Zheng et al., 2013; Liu et al., 2014; Paisitkriangkrai et al., 2015). It is found that spatial structure information of person's appearance is a very important cue, thus this information is integrated into the feature representation. For instance, (Farenzena et al., 2010) (known as SDALF) proposed to utilise symmetry and asymmetry in body structure to differentiate body parts from the background. Similar in spirit, pictorial structures (PS) (Cheng et al., 2011) are proposed based (Andriluka et al., 2009). Both methods attempt to suppress unnecessary background information focusing on the body parts during the feature matching, looking for part-to-part correspondences. Similar ideas are proposed: triangulated graphs in (Gheissari et al., 2006), uniform horizontal strip in (Gray and Tao, 2008; Layne et al., 2012; Prosser et al., 2010; Zheng et al., 2013; Liu et al., 2012), localised patches (Zhao et al., 2013c,b, 2014b; Liu et al., 2014; Li et al., 2014a; Bak et al., 2010; Zheng et al., 2015b; Paisitkriangkrai et al., 2015; Liao et al., 2015).

2. Model Learning (Unsupervised) After designing features, model learning can be performed. The aim of the model learning is to learn a function that projects the low-level (hand-crafted) feature representations into another subspace. This subspace is assumed to be more discriminative than original low-level feature space. That is, after the data represented by low-level features is projected into the subspace, the samples with the same identity are located closer than that of different identities, which is essentially clustering. How to learn this discriminative subspace in a unsupervised manner is the key to obtain a good performance. Towards this goal, there are several methods proposed for learning the subspace in different forms. Particularly, Wang et al. (2016a) proposed a semi-supervised kernel subspace learning model that learns cross-view identity-specific information from unlabelled data. Peng et al. (2016) proposed a novel method based on dictionary learning framework. Specifically, a multi-task dictionary learning model was formulated to transfer a view-invariant representation learned from a number of existing labelled source datasets to an unlabelled target dataset. The transferring was achieved a graph regularisation. Note that the methods (Wang et al. (2016a) and Peng et al. (2016)) used the conventional graph regularisation and rely on the graph constructed using hand-craft features which are prone to outlier and noise, whereas the thesis proposes a robust graph regularisation, and robust graph is learned with the dictionary jointly. Recently, unlike these two approaches that use features extracted from static images to learn discriminative features, to take advantage of image sequence information (multi-shot), Ma et al. (2017) proposed a new space-time person representation by encoding multiple granularities of spatio-temporal dynamics in the form of time series. This thesis does not focus on multi-shot person re-identification.

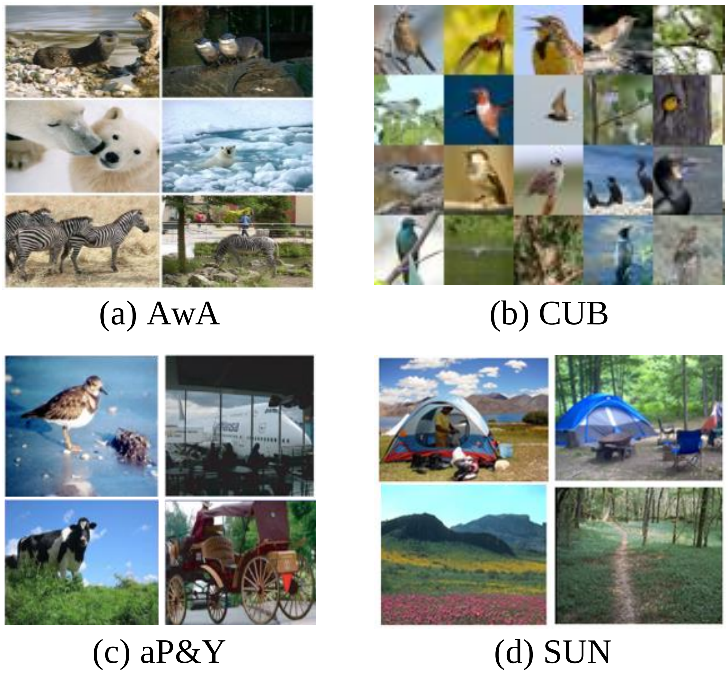
Unsupervised deep learning for ReID. Very recently deep learning methods for ReID have also been proposed in the unsupervised setting. These methods learn person discriminative features from raw visual data reducing the burden on feature engineering. In other words, the two components of ReID that are feature representation and model learning are performed under a single framework. Specifically, Hehe et al. (2017) proposed a progressive unsupervised learning (PUL) method to transfer pretrained deep representations to unseen domains. PUL mainly learns features by iterating two steps: 1) pedestrian clustering and 2) fine-tuning of the convolutional neural network to improve the original model trained on the irrelevant labelled dataset. Similarly, Schumann et al. (2017) formulates a deep learning based novel approach to automatic prototype-domain discovery for domain perceptive person re-identification. After that, a separate model for each of the discovered prototype-domains is learned, and during model testing, use the person probe image to automatically select the model of the closest prototype-domain. Overall, all methods including the proposed approach in this thesis have a common pattern in which they rely on clustering assumption e.g., similar persons in terms of features belong to the same class (ID). All these aforementioned approaches are generic that they may readily be applied to other kinds of visual verification tasks such as face/car verification (Cinbis et al., 2011).

Supervised ReID. Despite the success of usupervised CCTL methods for ReID, supervised learning methods are shown to be far better than unsupervised learning ones. When hand-crafted features and corresponding pairwise constraints are utilised, the optimal cross-view matching function is learned by either distance metric learning (Weinberger et al., 2005; Davis et al., 2007; Guillaumin et al., 2009), learning to rank (Prosser et al., 2010), or discriminant subspace learning (Xiong et al., 2014b). Recently these learning methods are incorporated into the supervised deep learning framework, achieving significant performance boost (Wei et al., 2017; Xiao et al., 2016; Li et al., 2014b; Yi et al., 2014; Ding et al., 2015). Specifically, a filter pairing neural network is particularly designed for jointly handling the misalignment, photometric and geometric transforms, occlusions and background clutter issues in ReID (Li et al., 2014b). Yi et al. (2014) applied a symmetric 'siamese' neural network to learn ReID features that are robust to the inherent challenging cross-view changes. More recently, a triplet-based ReID feature learning model is derived in (Ding et al., 2015; Hermans et al., 2017), where each triplet unit contains a query image from one view, a true and false match from another view.

Nevertheless, supervised model learning requires a large number of exhaustively labelled data. This assumption significantly limits their scalability in real-world scenarios. Therefore, although unsupervised learning is very challenging, its impact is significant for practical verification systems.

2.5 Benchmark Datasets

The proposed algorithms are evaluated on a number of benchmark visual datasets. They include object and action recognition benchmark datasets. Table 2.1 and Table 2.2 list the datasets. Examples corresponding to the datasets are shown in Figure 2.16, Figure 2.18, and Figure 2.17. The description of how each dataset is detailed in corresponding chapters.



(a) AwA

(b) CUB

(c) aP&Y

(d) SUN

Figure 2.16: Examples: (a) AwA , (b) CUB, (c) aP&Y, and (d) SUN datasets.

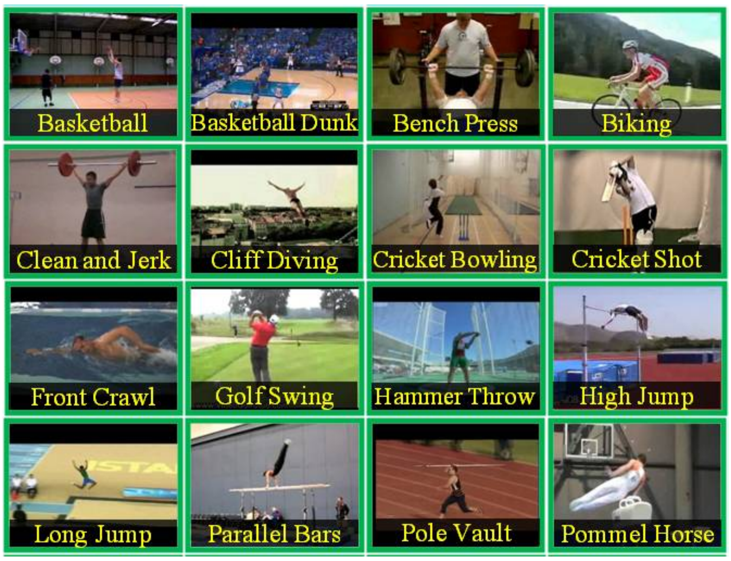


Figure 2.17: Examples from UCF101 dataset.

Dataset	Instances	SS	SS-D	Seen/Unseen	Chapter
AwA (Lampert et al., 2009)	30,475	A	85	40 / 10	3, 4
CUB (Wah et al., 2011)	11,788	A	312	150 / 50	3, 4
aP&Y (aPascal-aYahoo) (Farhadi et al., 2009)	15,339	A	64	20 / 12	3, 4
UCF101** (Soomro et al., 2012)	13,320 clips	A	115	51 / 50	3
SUN (Genevieve et al., 2014)	14,340	A	102	645 / 72 $^{(*)}$	4
ImageNet-1 (Russakovsky et al., 2015)	$1{,}2 \times 10^6$	W	1,000	800 / 200	4
ImageNet-2 (Russakovsky et al., 2015)	218,000	W	1,000	1,000 / 360	4

Table 2.1: Benchmark datasets for evaluation of transductive and inductive cross-class transfer learning methods. Notation: 'SS' – semantic space, 'SS-D' – the dimension of semantic space, 'A' – attribute, and 'W' – word vector. $^{(*)}$ – another split of 707/10 is also used for SUN (Jayaraman and Grauman, 2014; Zhang and Saligrama, 2016). Among datasets, aP&Y and SUN datasets provide instance-level atribute vectors, while the remaining ones are class-level. Following the literature, partitions for seen/unseen in all datasets are fixed. '**' indicates the action dataset.

Dataset	Cameras	Persons	Instances	Chapter
VIPeR (Gray et al., 2007)	2	632	1264	5
PRID (Hirzer et al., 2011)	2	749	949	5
CUHK01 (Li et al., 2012)	2	971	1,942	5
CUHK03 (Li et al., 2014b)	6	1,467	14,097	5

Table 2.2: Benchmark datasets for evaluation of unsupervised cross-class transfer learning method (person re-identification).

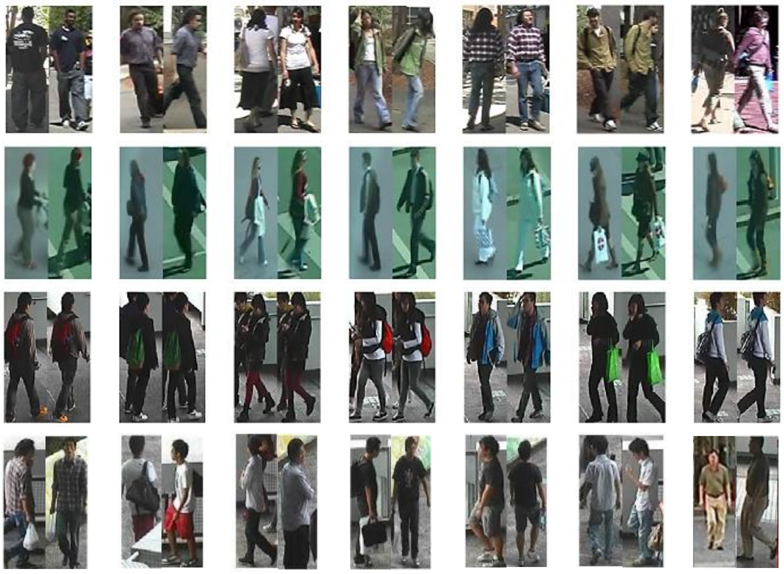


Figure 2.18: Examples: VIPeR (row-1), PRID (row-2), CUHK01 (row-3), and CUHK03 (row-4).

Chapter 3

Transductive Cross-class Transfer Learning by Unsupervised Domain Adaptation

Cross-class transfer learning (CCTL) requires to extract the knowledge from the labelled training dataset (source domain) and to transfer that knowledge to the test dataset (target domain). Since the target domain has no labelled data, existing CCTL methods adopt a naive transfer learning approach in which a model learned from the source domain is applied to the target domain blindly without any model adaptation. More specifically, existing CCTL methods typically assume that there is a semantic embedding space within which both the feature space and the class label spaces of the source and target domains can be embedded e.g., the class label ’cat’ can be represented as a binary attribute vector in attribute space, or as a high-dimensional word vector in word vector space. The embedded label vector in any given semantic space is called a class prototype. Given a semantic embedding space, most existing methods take a *visual feature embedding* approach. Specifically, the knowledge extracted from the source data is represented in the form of the embedding function that embeds each feature vector to its class prototype as an attribute or word vector. The embedding function can then be applied to the target image data to embed/project them into the same semantic space. After such embeddings, the classification of these target images can be simply nearest neighbour distance matching to the target class prototypes in the semantic space (see Section 2.2). Without adapting the learned embedding function to the target domain, existing methods are prone to the projection domain shift problem (see Figure 2.12).

In this chapter, in order to alleviate the domain shift problem, a novel method is proposed by developing a new unsupervised domain adaptation model under the transductive setting, thus transductive cross-class transfer learning [1] (TCCTL). TCCTL itself is not a unsupervised domain adaptation problem because the two domains have different tasks/classes. Note that the definition of the unsupervised domain adaptation (transductive transfer learning) in Pan and Yang (2010) is that two domains have the same task/classes. However, taking a visual feature embedding approach, the learning of embedding function for the target domain is a standard domain adaptation problem – both domains are embedded into the same semantic space (attribute or word vector) – albeit an unsupervised one as no label is available in the target domain. Uniquely, instead of learning a typical classification/regression function as in most previous works (Lampert et al., 2009; Frome et al., 2013; Fu et al., 2015b; Akata et al., 2015), learning of the embedding function is treated as a dictionary learning problem: Each dimension of the semantic embedding space corresponds to a dictionary basis vector and the coefficients/sparse code of each visual feature vector is a semantic representation in the semantic space. To learn a meaningful dictionary/embedding function for the target data, two important regularisation terms are introduced into the dictionary learning objective function, making a regularised dictionary learning model designed specifically for unsupervised domain adaptation. The first term controls the adaptation strength from the source domain to the target domain, whilst the second term rectifies explicitly the domain shift problem in CCTL, forcing the embedded target data to be near to the unseen class prototypes.

This chapter is organised as follows: Section 3.1 presents the details of the proposed approach for transductive cross-class transfer learning. Then, description of benchmark datasets, experimental settings, results with comparison to state-of-the-art methods, and ablation study are given in Section 3.2. Finally, a conclusion is presented in Section 3.3

3.1 Methodology

3.1.1 Problem Formulation

Suppose a source domain (a training set) is composed of c_s source classes with n_s labelled instances denoted as $\mathcal{D}_s = \{\mathbf{X}_s, \mathbf{S}_s, \mathbf{z}_s\}$, and similarly c_t target classes with n_t unlabelled instances for target domain (a test set) denoted as $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{S}_t, \mathbf{z}_t\}$. Each instance is represented using a

[1] Transductive cross-class transfer learning and transductive zero-shot recognition/classification terms are used interchangeably, since they have the same meaning in this thesis.

$d-$dimensional visual feature vector, that is, $\mathbf{X}_s = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_s}] \in \mathbb{R}^{n_s \times d}$ and $\mathbf{X}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_t}] \in \mathbb{R}^{n_t \times d}$, where $\mathbf{x}_i \in \mathbb{R}^d$. $\mathbf{z}_s \in \mathbb{R}^{n_s}$ and $\mathbf{z}_t \in \mathbb{R}^{n_t}$ are class label vectors for the source and target data respectively. It is assumed that the source and target classes are disjoint: $\mathbf{z}_s \cap \mathbf{z}_t = \varnothing$. Given a semantic space, $\mathbf{S}_s = [\mathbf{s}_1, \ldots, \mathbf{s}_{n_s}] \in \mathbb{R}^{n_s \times m}$ and $\mathbf{S}_t = [\mathbf{s}_1, \ldots, \mathbf{s}_{n_t}] \in \mathbb{R}^{n_t \times m}$ are the $m-$dimensional semantic representations, $\mathbf{s}_i \in \mathbb{R}^m$, of each class label in the source and target domains respectively (e.g., m-dimensional binary attribute vectors). For the source domain, $\mathbf{S}_s$ is given because each visual instance $\mathbf{x}_i$ of the source data is labelled by either a binary attribute vector or a word vector representing its corresponding class label z_s^i. In contrast, $\mathbf{S}_t$ has to be estimated because the target domain is unlabelled. The problem of transductive cross-class transfer learning (transductive zero-shot recognition) is thus to estimate $\mathbf{S}_t$ and $\mathbf{z}_t$ given $\mathbf{X}_t$, $\mathbf{X}_s$, and $\mathbf{S}_s$. Note that in the transductive setting, the unseen classes and the corresponding semantic prototypes need to be known in advance, but their equivalent images are not given.

3.1.2 Dictionary Learning for Embedding Function Learning

The aim is to learn an embedding function to map each d-dimensional visual feature vector $\mathbf{x}_i$ in $\mathbf{X}_s$ or $\mathbf{X}_t$ to a m-dimensional semantic embedding vector $\mathbf{s}_i$, i.e., $m < d$ a lower dimensional subspace is sought to embed $\mathbf{x}_i$ into. In the context of cross-class transfer learning (cross-class recognition), the subspace is the semantic space (attribute or word). In this work, the learning of the visual space to semantic space embedding is formulated as a dictionary learning and sparse coding problem. Sparse coding aims to represent a data vector as a sparse linear combination of basis elements, which are atoms of a learned dictionary. Taking attribute space as an example, to embed a data point from the visual feature space (higher dimensional) to an attribute space (lower dimensional), it is considered that each basis element (atom) corresponds to an attribute (or an axis in the attribute space). For example, to represent the attribute of ‘has fur’ in an animal image, a corresponding sparse coding coefficient of the image is the weight of that basis element in the image which represents how much fur is present in that image[2]. Denote the dictionary as $\mathbf{D} \in \mathbb{R}^{d \times m}$, a visual feature vector $\mathbf{x}_i$ can be reconstructed as $\mathbf{D}\,\mathbf{s}_i$ using its coefficient vector/projection $\mathbf{s}_i$ and the dictionary/embedding matrix $\mathbf{D}$. Dictionary learning is thus to learn $\mathbf{D}$ and $\mathbf{s}_i$ to minimise the reconstruction error. Since each dictionary basis has clear semantic meaning based on the definition given above, the learned dictionary is referred to as *semantic dictionary*.

[2] This is related to the concept of relative attributes (Parikh and Grauman, 2011).

Next, the dictionary learning problem is formulated separately for the source and target domains respectively, and highlight the difference in their formulations. First, in the source domain the sparse coding coefficient vector for each visual instance is known: For each $\mathbf{x}_i$, its corresponding $\mathbf{s}_i$ is the embedding (attribute or word vector) of its class label z_s^i in the semantic space. This is very different from the conventional dictionary learning by sparse coding whereby $\mathbf{s}_i$ needs to be estimated together with $\mathbf{D}$. Let us denote the source domain semantic dictionary as $\mathbf{D}_s$, the dictionary learning problem can be solved by quadratic optimisation:

$$\mathbf{D}_s = \min_{\mathbf{D}_s} ||\mathbf{X}_s - \mathbf{D}_s\mathbf{S}_s||_F^2, \;\; s.t. \;\; ||\mathbf{d}_i||_2^2 \leq 1, \tag{3.1}$$

where $||\cdot||_F$ is Frobenius norm of a matrix. It is a standard least squares minimisation problem with a closed-form solution. To avoid trivial solutions, a regularisation term is added to favour a solution of smaller norm. Eq. (3.1) thus becomes:

$$\mathbf{D}_s = \min_{\mathbf{D}_s} ||\mathbf{X}_s - \mathbf{D}_s\mathbf{S}_s||_F^2 + \lambda||\mathbf{D}_s||_F^2 \;\; s.t. \;\; ||\mathbf{d}_i||_2^2 \leq 1, \tag{3.2}$$

where λ controls the strength of the regularisation term. This is known as a least squares problem, also with a closed-form solution (Hoerl and Kennard., 1970).

Second, contrary to the source domain, the formulation for dictionary/embedding function learning by sparse coding in the target domain requires the conventional sparse coding mechanism as both the dictionary and the coefficient vectors are unknown and need to be learned from data:

$$\{\mathbf{D}_t, \mathbf{S}_t\} = \min_{\mathbf{D}_t, \mathbf{S}_t} ||\mathbf{X}_t - \mathbf{D}_t\mathbf{S}_t||_F^2 + \lambda||\mathbf{S}_t||_1 \;\; s.t. \;\; ||\mathbf{d}_i||_2^2 \leq 1, \tag{3.3}$$

where $||\mathbf{S}_t||_1 = \sum_{i=1}^{n_t} ||\mathbf{s}_i^t||_1$. In this formulation, the model is essentially learning a sparse representation of the data in an unsupervised fashion. Since both $\mathbf{D}_t$ and $\mathbf{S}_t$ are unknown and unconstrained (apart from enforcing S_t to be sparse), there is no guarantee that the learned representation captures a suitable semantic embedding space so that $\mathbf{D}_t$ is the correct embedding function for $\mathbf{X}_t$ and the embedding $\mathbf{S}_t$ in the semantic space is meaningful for CCTL. Therefore, learning $\mathbf{D}_t$ without any regularisation is undesirable meaning $\mathbf{D}_t$ has no use for CCTL.

Such a regularisation could come from the labelled source domain. Following the conventional naive transfer/non-adaptation CCTL approach, $\mathbf{D}_s$ is learned from the source domain (Eq. (3.1)) and then applied directly to the target data. This method, which forces $\mathbf{D}_t = \mathbf{D}_s$ rather than allowing $\mathbf{D}_t$ to be adapted from $\mathbf{D}_s$, is prone to the domain shift problem. To overcome this problem, both $\mathbf{D}_s$ and the target domain class prototypes are used to regularise the learning of $\mathbf{D}_t$.

This results in a novel unsupervised domain adaptation method for TCCTL based on regularised dictionary learning.

3.1.3 Domain Adaptation by Regularised Dictionary Learning

Now, it is time to introduce two critical regularisation terms into Eq. (3.3) to impose (a) *an adaptation regularisation constraint*: The $\mathbf{D}_t$ learned from the unlabelled target data should be similar to $\mathbf{D}_s$ learned from the labelled source data; and (b) *a visual-semantic similarity constraint*: The 'closeness' of the embeddings of target data ($\mathbf{s}_i$) to their true class labels in the semantic space (i.e., unseen class prototypes denoted as $\mathbf{p}^t_i$ and $i \in \{1,\dots,c_t\}$). This defines the new regularised dictionary learning framework:

$$\min_{\mathbf{D}_t,\mathbf{S}_t} \|\mathbf{X}_t - \mathbf{D}_t\mathbf{S}_t\|_F^2 + \lambda_1\|\mathbf{D}_t - \mathbf{D}_s\|_F^2 + \lambda_2\sum_{i,j} w_{ij}\|\mathbf{s}_i - \mathbf{p}^t_j\|_2^2 + \lambda_3\|\mathbf{S}_t\|_1 \ \ s.t. \ \ \|\mathbf{d}_i\|_2^2 \leq 1. \qquad (3.4)$$

Adaptation regularisation constraint (AR). Compared to Eq. (3.3), the new regularisation term $\|\mathbf{D}_t - \mathbf{D}_s\|_F^2$ in Eq. (3.4) regularises the amount of adaptation (closeness) of the learned dictionary $\mathbf{D}_t$ to the supervised learned dictionary $\mathbf{D}_s$. This term makes sure that the learned $\mathbf{D}_t$ is also a semantic dictionary that projects a target data point from the feature space to the same semantic space as $\mathbf{D}_s$. In this regard, $\mathbf{D}_s$ is treated as a basis for learning the dictionary $\mathbf{D}_t$ so that $\mathbf{D}_t$ is not deviated freely from $\mathbf{D}_s$. Without this regularisation, $\mathbf{D}_s$ could be adapted towards a trivial solution especially when $n_t > n_s$, i.e., the target data outnumbers the source data.

Visual-semantic similarity constraint (VSS). The second new regularisation term $\sum_{i,j} w_{ij}\|\mathbf{s}_i - \mathbf{p}^t_j\|_2^2$ in Eq. (3.4) enforces the visual-semantic similarity constraint. This is used to ensure that the learned coefficient vector $\mathbf{s}_i$ for each target data (its embedding in the semantic space) is close to its true class label z^t_i, embedded in the semantic space as $\mathbf{p}^t_i$. Since z^t_i is unknown, an estimate is obtained by visual-semantic similarity matching using the indirect attribute prediction (IAP) method (Lampert et al., 2009), where a probability is computed for $\mathbf{x}_i$ being labelled as z^t_j which defines the closeness of $\mathbf{s}_i$ to $\mathbf{p}^t_j$. Formally, the probability of $\mathbf{x}_i$ being the j-th unseen class is used as weight w_{ij} to enforce a closeness in the distance between the embedding/projection $\mathbf{s}_i$ and the j-th unseen class prototype $\mathbf{p}^t_j$, resulting in this regularisation term defined as $\sum_{i,j} w_{ij}\|\mathbf{s}_i - \mathbf{p}^t_j\|_2^2$. Note that this constraint utilises visual-semantic similarity matching whilst the dictionary aims to estimate the optimal visual feature embedding function. In the following, the proposed model is referred to as DTCCTL in which first letter 'D' stands for dictionary.

Discussion. Above $\mathbf{D}_t$ is regularised by $\mathbf{D}_s$, however one may ask whether the opposite possible.

It is common in literature that learning $\mathbf{D}_t$ subject to source domain knowledge is desirable, because the ultimate goal is to perform well in the target domain Fernando et al. (2013) not in the source domain.

3.1.4 Optimisation

It is important to point out that Eq. (3.4) is not convex for $\mathbf{D}_t$ and $\mathbf{S}_t$ simultaneously, although it is convex for each of them separately (biconvex). Thus an alternating optimisation method is deployed to solve it. In particular, the following two subproblems are solved alternatingly:

1. Fix $\mathbf{S}_t$, update $\mathbf{D}_t$. $\mathbf{D}_s$ is computed using Eq. (3.1), and $\mathbf{S}_t$ is initialised randomly.

$$\mathbf{D}_t^* = \arg\min_{\mathbf{D}_t} ||\mathbf{X}_t - \mathbf{D}_t\mathbf{S}_t||_F^2 + \lambda_1 ||\mathbf{D}_t - \mathbf{D}_s||_F^2 \tag{3.5}$$

This is a standard least squares problem, having a closed-form solution:

$$\mathbf{D}_t^* = (\mathbf{X}_t\mathbf{S}_t^T + \lambda_1\mathbf{D}_s)(\mathbf{S}_t\mathbf{S}_t^T + \lambda_1\mathbf{I})^{-1}. \tag{3.6}$$

2. Fix $\mathbf{D}_t$, update $\mathbf{S}_t$

$$\mathbf{S}_t^* = \arg\min_{\mathbf{S}_t} ||\mathbf{X}_t - \mathbf{D}_t\mathbf{S}_t||_F^2 + \lambda_2 \sum_{i,j} w_{ij}||\mathbf{s}_i - \mathbf{p}_j^t||_2^2 + \lambda_3||\mathbf{S}_t||_1 \tag{3.7}$$

In this equation, the first two terms can be combined into a single quadratic form and it becomes a conventional sparse coding problem. To solve it Lasso (Tibshirani, 1996) solver from the SPAMS toolbox (Mairal et al., 2010) is used.

The iterations will terminate when the objective function in Eq. (3.4) converges or after a fixed number of iteration. Note that a positive constraint is set on coefficients if the attribute space is used because it does not make sense to have a negative attribute value. For semantic word space, this constraint is removed.

3.1.5 Cross-class Recognition

Single Semantic Space: Once the dictionary coefficients $\mathbf{S}_t$ is estimated, cross-class recognition can be performed. In this work, two classification strategies are considered: 1) a nearest neighbour (NN) classifier (see Section 2.2) and 2) a semi-supervised label propagation (LP) framework. For the NN classifier, given a target data $\mathbf{x}_i$, its coefficients $\mathbf{s}_t^i$ is directly used to compare with the unseen prototypes. It is then labelled as the nearest unseen class. For the LP classifier, the method in (Fu et al., 2014) is adopted. Specifically, the unseen data and the unseen prototypes (as the labelled data) are exploited to set up a graph, then the label information is propagated from the unseen prototypes to each unseen data. The performance of the proposed algorithm (DTC-CTL) on both strategies are reported.

Combining Multiple Spaces: Multiple semantic spaces can be easily combined in the proposed framework to exploit their complementarity. For example, after estimating $\mathbf{S}_t^A$ and $\mathbf{S}_t^W$ for attribute and word space, respectively, the similarity matrices from these two spaces can be combined by a simple strategy: For the NN classifier, the distances to neighbours are averaged; for the LP classifier, the graph similarity matrix are averaged before label propagation.

3.2 Experiments and Evaluations

3.2.1 Datasets and Settings

Datasets. Four datasets are used in the experiments. (a) *AwA* (Lampert et al., 2009) consists of 30,475 animal images belonging to 50 classes. An 85D attribute vector is provided for each class. (b) *CUB* (Wah et al., 2011) is a fine-grained dataset, containing 200 different bird classes, with 11,788 images in total. The class level attribute annotations are given with 312 visual attributes (e.g., color, part pattern). (c) *aPascal-aYahoo* (Farhadi et al., 2009) consists of two datasets: aPascal is a 12,695 images subset of the PASCAL VOC 2008 dataset and aYahoo has 2,644 images. A 64D attribute vector is provided for each image. There are 20 object classes for aPascal, and 12 for aYahoo and they are disjoint. aPascal-aYahoo is used for cross-dataset cross-class transfer learning (see Section 3.2.4). (d) *UCF101* (Soomro et al., 2012) is one of the largest datasets for action recognition with 101 classes, containing 13,320 video clips and 27 hours of video data in total. The videos are collected from YouTube with large camera motions and cluttered background making them particularly challenging. A 115D attribute vector is provided for each action e.g., body posture, body part motion. The summary of these datasets is given in Table 2.1.

Features. Two types of features are used: deep features (CNN) and hand-crafted features (L). For the AwA, OverFeat is used to extract 4,096D feature vectors (Sermanet et al., 2013). For the hand-crafted features, publicly available features are used, the same as in previous work (Lampert et al., 2009; Akata et al., 2013; Fu et al., 2014). For CUB, deep features are extracted the same as

AwA. Since hand-crafted features are not provided with the dataset, based on the setting of (Akata et al., 2013), 96 color descriptors are extracted from regular grids at three scales, and aggregate them into fisher vectors (FVs) using 256 Gaussians. For aPascal-aYahoo, 9,751D hand-crafted features provided by (Farhadi et al., 2009) is used. For the UCF101, the features provided by the THUMOS challenge (Jiang et al., 2014) which contain 4,000D Motion Boundary Histogram (MBH) features (Wang and Schmid, 2013) are used.

Settings. Two types of semantic embedding spaces are used: (a) An attribute space where each class is represented as a binary attribute vector. (b) An 100D semantic word space learned by the skip-gram model (Mikolov et al., 2013) using a text corpus containing 4.6M Wikipedia pages. For the source/target class split, the standard 40/10 split is used for AwA, while 150/50 split for the CUB as in (Akata et al., 2013). For aPascal-aYahoo, standard split is used with 20/12, in that, aPascal (20) is used for training, and aYahoo for testing (12). For the UCF101, two types of split are used: 81/20 and 51/50. In each experiment, the average recognition accuracy with standard errors is reported over 10 trials with different random splits.

There are four parameters in Eq. (3.4) and Eq. (3.2): λ λ_1, λ_2, and λ_3. It is found that λ and λ_3 have little influence on performance, and they thus set to 0.05 and 0.05 respectively. Other parameters are set by 5-fold cross validation technique with training data. In detail, A and W are attribute space and word space respectively: $\lambda_1^A = 0.001$, $\lambda_2^A = 0.005$; $\lambda_1^W = 0.03$, $\lambda_2^W = 0.005$ for AwA; $\lambda_1^A = 0.05$, $\lambda_2^A = 0.001$; $\lambda_1^W = 0.01$, $\lambda_2^W = 0.02$ for CUB; $\lambda_1 = 0.02$ $\lambda_2 = 0.005$ for aPascal-aYahoo; $\lambda_1^A = 0.05$ $\lambda_2^A = 0.001$; $\lambda_1^W = 0.5$ $\lambda_2^W = 0.01$ for UCF101. LP parameters are set empirically: the number of neighbours to construct similarity is 5, the parameter for balancing the propagation rate is 0.8 Fu et al. (2014). For the zero-shot classifier (see Section 3.1.5), LP is used for the reported results unless stated otherwise. More detailed analysis are left for the future work.

3.2.2 Comparative evaluation

Comparative models. For AwA, 11 most recent and competitive CCTL methods are selected for comparison, as shown in Table 3.1.

- **IAP (Lampert et al., 2009)** is probabilistic approach in which it assumes that all attributes are independent and equally important for learning the cross-class recognition. The connection between the seen and unseen classes is established indirectly by attributes as mid-level layers (see more details in C). Note that IAP cannot utilise semantic word vector, since it treats every attribue individually.

- **DAP (Lampert et al., 2009)** is also probabilistic approach, similar to IAP. The only difference is establishing the connection between the seen and unseen classes wherein all classes are treated equally based on the attributes' results in the attribute layer. Similar to IAP, DAP also cannot use semantic word vector.

- **DS (Rohrbach et al., 2010)** performs knowledge transfer by representing unseen object classes relative to known ones. More specifically, DS learns a model based on similarities between an unseen object class and known object classes. In DS, semantic word or attribute vector can be used to measure semantic relatedness between known and unseen classes.

- **AHLE (Akata et al., 2013)** views cross-class recognition as a label-embedding problem, that is, each class is embedded in the space of attribute vectors. To achieve this, on max-margin ranking loss is proposed which measures the compatibility between an image and a label embedding. This model can use multiple semantic information such as semantic attribute vector, word vector, and class hierarchies.

- **HEX (Deng et al., 2014)** is based on graph paradigm. Specifically, Hierarchy and Exclusion (HEX) graphs is introduced that captures semantic relations between any two labels applied to the same object including mutual exclusion, overlap and subsumption. For this method, only semantic attributes are used to build graphs.

- **TMV-BLP (Fu et al., 2014)** is transductive approach which is closely realted the appraoch proposed in this chapter. This method can use both semantic word vector and attribute vector (see Section 2.2 for details).

- **Yu et al. (2013)** proposed a novel formulation to automatically design 'category-level attributes', which can be encoded by a category-attribute matrix. Then, they proposed a framework of using attributes as mid-level cues for multi-class recognition on seen classes. Finally, the error of such recognition scheme is used to measure the discriminativeness of attribtes. Only attribute vectors are used for this model, while other semantic embeddings is difficult to incorporate.

- **Jayaraman and Grauman (2014)** proposed a novel random forest approach to train cross-class recognition models that explicitly takes care of the unreliability of attribute predicitons. Unrealibility is measured by statistics about each attribute's error tendencies – the attribute classifiers receiver operating characteristics. This approach uses attributes as semantic information.

- **CNNSVM (Ozeki and Okatani, 2014)** proposed a simple method: first features are learned with a deep convolutional neural network (CNN), and after learning the network, it is used a feature extractor. Finally, attribute classifiers are built by a linear SVM on top of the features extracted from the network. Only attributes are used for cross-class recognition.

- **AMP (Fu et al., 2015b)** is a method that models the semantic manifold in an embedding space using a semantic class label graph. The semantic manifold graph is used to redefine the distance metric in the semantic embedding space for more effective cross-class recognition. The semantic manifold distance is computed using a observing Markow chain process (AMP). This model can use both semantic word vector and attribute vector.

- **SJE (Akata et al., 2015)** is a method that uses structured joint embedding. SJE relates input data such as image features and output embeddings such as semantic information through a compatibility function, thus accounting for a structure in the output sapce. This model uses atributes, word vector and WordNet hierarchy for cross-class recognition.

These 11 models can be divided into several groups depending on various aspects: (1) Features: Most reported results on the dataset-provided hand-crafted features, although more recently the deep features have been used CNNSVM (Ozeki and Okatani, 2014), HEX (Deng et al., 2014), AMP (Fu et al., 2015b), SJE (Akata et al., 2015); (2) Side information (SI): This refers to what semantic information extracted from human knowledge is used. In addition to embedding each class label into either an attribute space (A) or word vector space (W), the Wordnet hierarchy (H) is used in AHLE (Akata et al., 2013) and SJE (Akata et al., 2015). Some methods such as (Yu et al., 2013; Jayaraman and Grauman, 2014) also use a different form of annotation, learned from the given category-attribute matrix, that is, instead of class attribute annotation, a class similarity (CS) matrix is used; (3) Most of them are based on the visual feature embedding approach, with the only exception of IAP (Lampert et al., 2009) which uses visual-semantic similarity matching. Mainly, the proposed approach (DTCCTL) in this chapter is very different from all the comparative models: (1) DTCCTL is based on dictionary learning framework, and (2) the paradigm of unsupervised domain adaptation is used. Also, DTCCTL is very flexible: can utilise unlabelled target data, similar to TMV-BLP; can use whatever semantic information if available simiarl to SJE , and use different features.

In contrast, far fewer studies have been reported on the more challenging CUB dataset (more and fine-grained classes). For the UCF101 action recognition dataset, no results have been reported so far, although Liu et al. (2011a) has tackled a similar CCTL problem for action recognition, albeit using different (much smaller) datasets. The method in (Liu et al., 2011a) is essentially based on learning embedding function to the attribute space using the source data and then using the embedding function to embed the target data in the attribute space followed by nearest neighbour-based classification. In addition to (Liu et al., 2011a), to obtain the results on DAP and IAP, the code in (Akata et al., 2013) is used. For all the competitors, kernelised SVMs are used to learn the attribute classifiers (embedding function to the attribute space, DAP) and source class classifiers (for IAP). In contrast, in the proposed model (DTCCTL), the learned dictionary acts as attributes classifiers. The use of the kernelised SVM for the attribute and class classifiers are as follows: say there are k number of attributes, and SVM is used to learn k number of attribute classifiers for each attribute; n-way class classfiers are learned for each class.

Performance Comparison.

AwA and CUB benchmarking – Table 3.1 shows that overall the proposed method (DTCCTL) has the best performance on these two image datasets. In particular, it is observed that: (1) On AwA, if the same hand-crafted features are used, DTCCTL's result (49.7%) is the highest; (2) The results of TMV-BLP give the most competitive alternative to DTCCTL in this setting. As discussed in Section 2.2, TMV-BLP and DTCCTL are the two which aim to rectify the projection domain shift problem by utilising the unlabelled target domain data. These results suggest that the proposed new regularised dictionary learning-based formulation is more effective than the two-step (projection followed by adaptation) approach taken by TMV-BLP; Note that the proposed approach in this thesis is different from the two-step approach: the proposed method does adaptation and learns a mapping function at the same time, while TMV-BLP projects all the data into some space first, and then perform adaptation; (3) When the more powerful deep features are used, DTCCTL gains a significant performance boost, rising from 49.7% to 75.6% and the

Method	F	SI	AwA	CUB
IAP (Lampert et al., 2009)	L/C	A	42.2/44.5	5.60/19.5
DAP (Lampert et al., 2009)	L/C	A	41.4/53.2	10.5/31.4
DS (Rohrbach et al., 2010)	L/C	W/A	35.7/52.7	-
AHLE (Akata et al., 2013)	L	A+H	43.5	18.0
HEX (Deng et al., 2014)	L/C	A	38.5/44.2	-
TMV-BLP (Fu et al., 2014)	L	A+W	47.1	-
Yu et al. (2013)	L	CS	48.3	-
Jayaraman and Grauman (2014)	L	A+CS	48.7	-
CNNSVM (Ozeki and Okatani, 2014)	C	A	62.4	-
AMP (Fu et al., 2015b)	C	A+W	66.0	-
SJE (Akata et al., 2015)	L/L+C	A+W+H	42.3/67.8	19.0/**47.1**
DTCCTL	L/C	A	47.5/73.2	26.7/ 39.5
DTCCTL	L/C	A+W	**49.7/75.6**	**28.1**/ 40.6

Table 3.1: Cross-class recognition results on AwA and CUB on the target domain (%). Notations – ’F’: features; ’L’: hand-crafted features; ’SI’: side information; ’C’: deep features; ’A’: attribute space; ’W’: semantic word vector space; ’H’: WordNet hierarchy; ’CS’: class similarity. When two results are reported, they correspond to the two types of features used.

gap to the best alternative (67.8% SJE) becomes bigger; (4) The same conclusion can be drawn on the CUB – DTCCTL’s overall results are superior to the compared methods. Note that SJE obtained better result using the deep features (47.1% vs. 40.6%) but its result on hand-crafted feature is much weaker than DTCCTL (19.0% vs. 28.1%). It is worth pointing out that SJE employ combined hand-crafted features and deep features, and use more than two semantic spaces. In contrast, other methods including the proposed model use only one type of features and no more than two semantic spaces. Richer features and more complementary semantic spaces would certainly help the proposed method as well but were not used to be fair to other compared methods. Analysing combining different features and trying more semantic spaces are left for future work.
UCF101 benchmarking – For this dataset, the results are shown in Table 3.2. Comparing Table 3.2 with Table 3.1, it is apparent that CCTL for action recognition from videos is a much harder task than object recognition on images. In particular, with 50 target classes in both CUB and

Method	SI	51/50 (%)	81/20 (%)
DAP	A	02.2 ± 0.5	06.1 ± 1.5
IAP	A	06.9 ± 1.1	11.1 ± 1.9
Liu et al. (2011a)	A	02.5 ± 1.2	06.2 ± 2.1
DTCCTL	A	13.2 ± 1.9	20.1 ± 3.1
DTCCTL	A+W	**14.0** ± **1.8**	**22.5** ± **3.5**

Table 3.2: Transcductive cross-class recognition results on the UCF101 dataset.

UCF101, the same DAP and IAP methods yielded much poorer results, close to the chance level (2%) in the case of DAP. In addition, the following observations can be made: (1) DTCCTL performs much better than the three compared alternatives, almost doubling the recognition rates of the best competitor (IAP) under both settings. Note that although an additional semantic embedding space (word space) is used, DTCCTL's results with attributes alone is still much better; (2) The very poor results from both DAP and Liu et al. (2011a) suggest that embedding without adaptation fails completely on this dataset. Moreover, it also suggests that using the source data to learn a n-way classifier for measuring the visual similarity is more sensible for video actions given the larger domain shift problem at hand. This explains the better performance of IAP than DAP and Liu et al. (2011a).

3.2.3 Comparison with Unsupervised Domain Adaptation Methods

In this experiment, it is demonstrated that unsupervised domain adaptation helps cross-class transfer learning, and the proposed regularised dictionary learning-based adaptation is better than the alternatives.

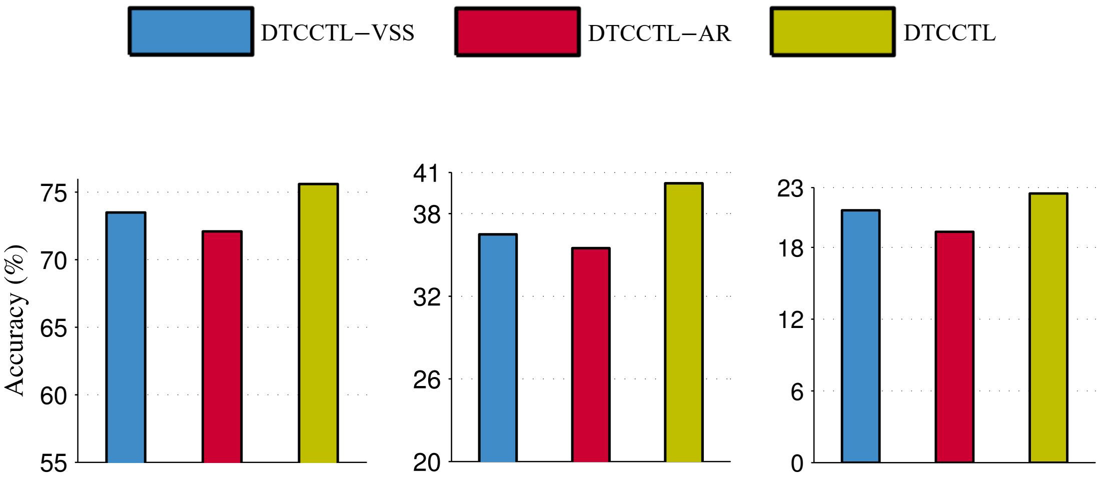
Competitors. For all three datasets, DTCCTL is compared with three most recent and relevant subspace alignment-based unsupervised domain adaptation methods: 1) Geodesic flow kernel (GFK) (Gong et al., 2012), 2) Subspace alignment domain adaptation (SADA) (Fernando et al., 2013), and 3) Subspace interpolation dictionary learning (SIDL) (Ni et al., 2013). All three methods attempt to align the data distributions of the two domains. When applied to the TCCTL problem, the projection function (based on the same dictionary learning model) learned in the source domain can thus be used directly for the target domain after they are aligned. In addi-

Method	AwA (%)	CUB (%)	UCF101(%)
GFK (Gong et al., 2012)	65.2	31.7	16.3
SIDL (Ni et al., 2013)	64.3	33.2	18.7
SADA (Fernando et al., 2013) 65.7	31.4	17.4	
DTCCTL-no-adapt	62.1	34.5	18.1
DTCCTL	**75.6**	**40.6**	**22.5**

Table 3.3: Evaluations on unsupervised domain adaptation methods. Deep features are used.

tion, the DTCCTL is also evaluated without adaptation constraint, that is, setting $\mathbf{D}_t = \mathbf{D}_s$ (see Section 3.1.2), denoted as DTCCTL-no-adapt.

Performance Comparison. Table 3.3 shows the comparative results. It can be seen that adaptation certainly helps in the DTCCTL: comparing DTCCTL with DTCCTL-no-adapt, a clear improvement can be observed thanks to the adaptation of $\mathbf{D}_s$ to $\mathbf{D}_t$ using Eq. (3.4). The results also show that the alternative subspace alignment-based adaptation methods are much weaker than DTCCTL. The results are slightly better than that without adaptation on AwA; but on the more challenging CUB dataset, their adaptations have an adverse effect. These results thus suggest that existing unsupervised domain adaptation methods are not effective under the CCTL setting. This is because that they are designed for visual recognition problems where each data can only have a single class label. In a multi-label scenario such as cross-class transfer learning (e.g., each AwA image can have dozens of attributes present), the subspace alignment strategy would not be a good strategy. This is particularly true for CUB where all images contain a bird and aligning the distributions of two sets of bird images will have little effect because the two distributions may have already been similar. The alignment thus would not help to solve the more subtle domain shift problem that the beak of a seagull is different from that of a pigeon. In contrast, the DTCCTL utilises the unseen class prototypes to regularise the learning of target domain embedding function which is specifically designed for rectifying the domain shift problem for cross-class transfer learning.



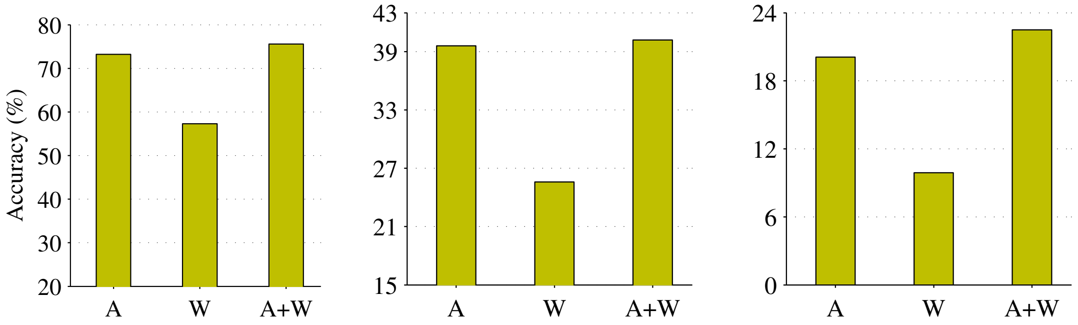
(a) AwA (b) CUB (c) UCF101

Figure 3.1: Evaluation of the contributions of each component of the DTCCTL on AwA, CUB and UCF101 (deep features).

3.2.4 Further Analysis

Ablation Study. In Figure 3.1, the proposed full model (DTCCTL) is compared with various stripped-down versions of the model to validate the contributions of the two regularisation terms in Eq. (3.4). Specifically DTCCTL (Eq. (3.4)) is compared with DTCCTL without the visual-semantic similarity constraint, i.e., Eq. (3.4) without the regularisation term $\sum_{ij} w_{ij} ||\mathbf{s}_i - \mathbf{p}_j^t||_2^2$ (denoted DTCCTL–VSS, '–' for minus) and DTCCTL without the adaptation regularisation constraint (DTCCTL–AR). The results in Figure 3.1 show clearly that both regularisation terms contribute to the superior performance of DTCCTL.

Effects of Combining Multiple Semantic Spaces. In the proposed model, the attribute and semantic word vector space are combined in the label propagation-based cross-class recognition algorithm (see Section 3.1.5). Figure 3.2 shows the results of the DTCCTL when only one of the two semantic embedding spaces is used. It can be seen that the model performance is notably improved by utilising both semantic spaces. It is also noted that using just one semantic space, the model already achieves very competitive performance. Moreover, the performance in the attribute space is stronger compared to the word vector space because the latter is unsupervised and does not benefit from human annotation. In particular, it is observed that the semantic word space is much weaker for UCF101. This is because the action names such as 'apply lipstick'



(a) AwA (b) CUB (c) UCF101

Figure 3.2: Effectiveness of combining multiple semantic embedding spaces (deep features are used).

and 'ski-jet' are much more abstract and ambiguous to describe the rich content of the associated actions, compared to the nouns in the image datasets (e.g., 'giant panda'). Simply embedding the class names to the semantic word space may not be the best way to explore the word space for cross-class (zero-shot) action recognition especially for subtle and complex actions.

Effects of the classification methods. The results reported so far are obtained using the label propagation (LP) classifier after domain adaptation. Table 3.4 shows that when the nearest neighbour classifier (NN) with cosine distance is used the performance is only slightly worse, by about 2%.

	AwA(%)	CUB(%)	UCF101(%)
NN	74.1	38.4	20.1
LP	**75.6**	**40.2**	**22.5**

Table 3.4: Classification methods: nearest neighbour (NN) vs. label propagation (LP).

The effects of the amount of target data used. One of the key differences between the proposed model and the alternatives except (Fu et al., 2014) is that unlabelled target data is used for the model learning. This is determined by the nature of the proposed approach – a transfer learning method with any form of adaptation to the target data needs to use the target data. In this experiment, it is evaluated that how the learned embedding function is affected by the amount of target data used in the model learning. Figure 3.3 suggests that the impact is very small. The performance on all three datasets only drops slightly with as few as about 2% of the target data. Furthermore, it is important to note that when no target data is used, that is, $\mathbf{D}_t = \mathbf{D}_s$, the dictio-

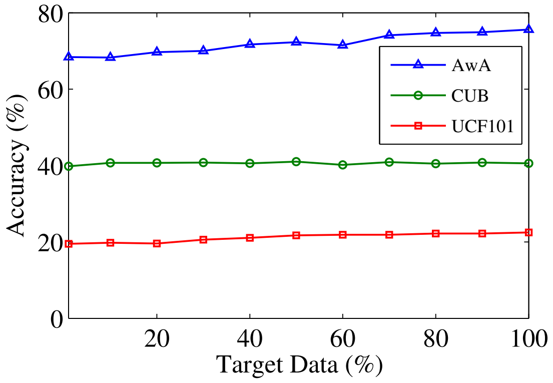


Figure 3.3: The effect of the amount of target data used.

Method	aP&Y (%)
IAP	16.9
DAP	16.8
DTCCTL	**26.5**

Table 3.5: Evaluations on cross-dataset cross-class transfer learning.

nary learning just by using source data itself gives very competative performance compared to using the target data showing the dictionary learning is indeed a very promising direction. However, to achieve the full performance the proposed model need to use all target data, this is due to the transductive setting of the proposed model as mentioned in the introduction section.

Cross-dataset cross-class transfer learning. In this experiment, by following the same setting as in (Lampert et al., 2009), the proposed model is evaluated on using aPascal as source data and aYahoo as target data (Farhadi et al., 2009). Since these datasets have per-image attribute annotations dictionary is learned with per-image labels. Cross-class recognition accuracy of 26.5% is obtained by the proposed model with NN as classifier, while with exactly the same features, the DAP and IAP results in (Lampert et al., 2009) are 16.8% and 16.9% respectively - about 10% lower than the proposed model (see Table 3.5).

Attribute prediction with and without domain adaptation. The attribute prediction accuracy on the target data of the proposed model with and without domain adaptation is evaluated using the AUC metrics as in (Lampert et al., 2009). Without domain adaptation, the results on AwA, CUB, aPascal-aYahoo are 65.5%, 54.1%, and 56.7%, respectively, whereas the results with do-

Method	AwA (%)	CUB (%)	aP&Y (%)
DTCCTL-no-adapt	65.5	54.1	56.7
DTCCTL	69.1	57.8	59.2

Table 3.6: Attribute prediction with and without domain adaptation.

main adaptation are 69.1%, 57.8%, and 59.2%, respectively (see Table 3.6). This suggests that domain adaptation leads to better attribute prediction accuracy, which in turn contributes to the better CCTL performance.

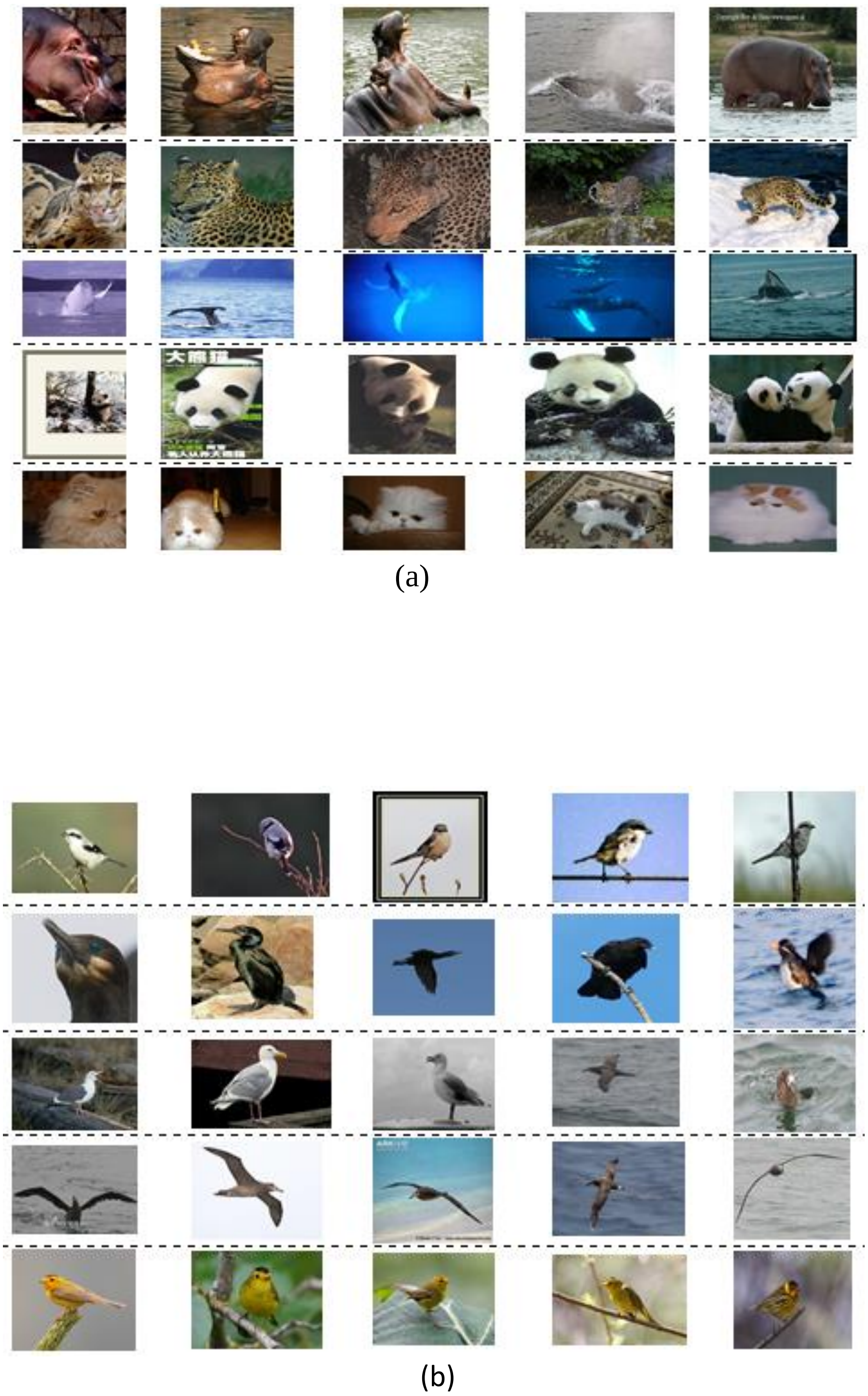
Dictionary visualisation. Figure 3.4 shows visualisation of the learned dictionary on AwA and CUB datasets. The dictionary basis is visualised via exemplar images from dataset. Each row in the Figure shows five example images whose features are closest to a certain basis (a column vector in the learned dictionary). Specifically, d-dimensional feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ are extracted for all given images. Then, a basis is chosen which is also a d-dimensional vector $\mathbf{d}_i \in \mathbb{R}^d$. After that, the distance between the feature vectors and the basis is computed, and top 5 nearest feature vectors are chosen in terms of distance. The exemplars suggest that the learned basis capture certain common properties across multiple categories. For example, the basis in row-1 may be described by the 'brown colour' texture pattern, row-2 may be viewed close to concepts of 'dotted and striped' pattern, row-3 may be described by the 'blue texture' pattern. Similarly, certain concepts can be seen in the CUB dataset.

Qualitative results on UCF101 dataset. Some qualitative results are given in Figure 3.5. The proposed model is compared to DAP, IAP, and Liu et al. (2011a) methods.

3.3 Conclusion

This chapter proposed a novel cross-class transfer learning framework under transductive setting. The framework is build on dictionary learning for sparse coding with two novel critical regularisations. The regularisations are shown to play a major role in which they enable the model learn an embedding function from a visual space to a semantic space using labelled source, unlabelled target data, and target class prototypes. Therefore, compared with most existing cross-class transfer learning methods that perform naive transfer, the proposed model can be considered as essentially an unsupervised domain adaptation model for cross-class transfer learning.

Extensive experiments are conducted on action and image benchmark datasets with both



(a)

(b)

Figure 3.4: Semantic dictionary visualisation. Example images whose features are closest to a certain basis. Each row corresponds to a certain dictionary basis.

Action	DTCCTL	DAP	IAP	Liu et al.
‘Drumming’ 	**‘Drumming’** ‘Playing Flute’ ‘Frisbee Catch’ ‘Uneven Bars’ ‘Skiing’	‘Mopping Floor’ **‘Drumming’** ‘Punch’ ‘Frisbee Catch’ ‘Pull Ups’	‘Frisbee catch’ ‘Salsa spins’ ‘Jumping jack’ ‘Floor gymnastics’ ‘Uneven bars’	‘Frisbee Catch’ ‘Salsa Spins’ ‘Floor Gymnastics’ ‘Uneven Bars’ ‘Jumping Jack’
‘Apply lipstick’ 	**‘Apply Lipstick’** ‘Blow Dry Hair’ ‘Playing Flute’ ‘Blowing Candles’ ‘Playing Tabla’	‘Mopping Floor’ ‘Drumming’ ‘Punch’ ‘Frisbee Catch’ ‘Pull Ups’	‘Salsa spins’ ‘Frisbee Catch’ ‘Floor Gymnastics’ ‘Jumping Jack’ ‘Uneven Bars’	‘Salsa Spins’ ‘Frisbee Catch’ ‘Jumping Jack’ ‘Floor Gymnastics’ ‘Pull Ups’

Figure 3.5: Qualitative results on UCF101 dataset. For each example, the top-5 unseen predictions of the DTCCTL and the Euclidean distance are shown. The predictions are ordered by decreasing score, with the correct predictions in bold.

hand-crafted features and deep features. Firstly, it is found that performing cross-class recognition for action is more challenging than image-based cross-class recognition, due to complex nature of actions requiring more research. Secondly, the experimental results demonstrated that the proposed method is superior over existing methods. This is indeed attributed with the proposed regularisations which enable domain adaptation, and making use of the dictionary learning for sparse coding. Thirdly, it is found that the dictionary learning framework itself is promising. This is because the dictionary learning can achieve very competitive performance compared to state-of-the-art methods without any regularisation and unlabelled target data. Also, visualisation of the learned dictionary with the proposed model shows that dictionary atoms capture specific semantic information across a variety of classes. Finally, some qualitative results validate superior performance of the proposed approach.

Chapter 4

Inductive Cross-class Transfer Learning by Semantic Autoencoder

As discussed in the previous chapter, existing cross-class transfer learning models mostly suffer from the projection domain shift problem. In order to overcome this problem, transductive cross-class transfer learning approaches including the approach presented in Chapter 3 were shown to be promising. Their assumption is that unlabelled test data and test prototypes (without correspondence) are available for model learning. However, this assumption is often invalid in the context of cross-class transfer learning because new classes typically appear dynamically and unavailable before model learning. In this chapter, instead of assuming availability of all test unseen class data for transductive learning, a novel method is proposed based on inductive learning, thus *inductive cross-class transfer learning*[1]. Specifically, the novel approach is proposed to inductive cross-class transfer learning based on the encoder-decoder paradigm. An encoder projects a visual feature representation of an image into a semantic space such as an attributes space, similar to a conventional cross-class transfer learning model. However, the visual feature embedding as an input to a decoder is also considered in which the decoder aims to reconstruct the original visual feature representation. This additional reconstruction task imposes a new constraint in learning the visual → semantic embedding function so that the embedding must also preserve all the information contained in the original visual features, i.e. they can be recovered by the decoder. It is shown that this additional constraint is very effective in mitigating the domain

[1] Cross-class transfer learning, inductive cross-class transfer learning, and zero-shot recognition/classification are used interchangeably.

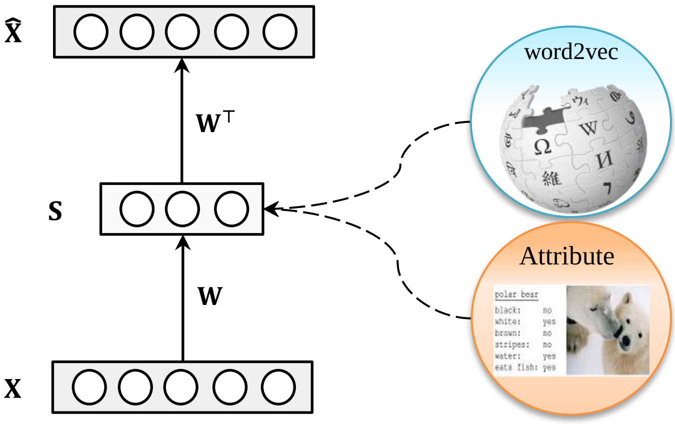


Figure 4.1: The proposed semantic autoencoder (SAE) leverages the semantic side information such as attributes and word vector, while learning an encoder and a decoder.

shift problem. This is because although the visual appearance of attributes may change from seen classes to unseen classes, the demand for more truthful reconstruction of the visual features is generalisable across seen and unseen domains, resulting in the learned embedding function less susceptible to the domain shift.

Mathematically, a semantic autoencoder (SAE) is formulated with the simplest possible encoder and decoder model architecture (see Figure 4.1): Both have one linear projection to or from a shared latent embedding/code layer, and the encoder and decoder are symmetric so that they can be represented by the same set of parameters. Such a design choice is motivated by computational efficiency – the true potential of a cross-class transfer learning model is when applied to large-scale visual recognition tasks where computational speed is essential. Even with this simple formulation, solving the resultant optimisation problem efficiently is not trivial. In this work, one such solver is developed whose complexity is independent of the training data size therefore suitable for large-scale problems. Notably, the proposed semantic autoencoder differs from conventional autoencoder the latent layer has clear semantic meaning: It corresponds to the semantic space and is subject to strong supervision. Therefore the proposed model (SAE) is not unsupervised.

This chapter is organised as follows. The detailed explanation of the proposed approach is presented in Section 4.1. This is followed by experiments with comprehensive evaluations with comparison to various state-of-the-art inductive cross-class transfer learning methods on both mid-scale and large-scale benchmark datasets in Section 4.2. Finally, Section 4.3 concludes this chapter.

4.1 Methodology

4.1.1 Linear Autoencoder

First the formulation of a linear autoencoder (AE) is introduced for completeness, and then extending AE into a semantic one is presented. In its simplest form, an autoencoder is linear and only has one hidden layer shared by the encoder and decoder. The encoder embeds the input data into the hidden layer with a lower dimension and the decoder projects it back to the original feature space and aims to faithfully reconstruct the input data. Formally, given an input data matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$ composed of N feature vectors of d dimensions as its columns, it is projected into a k-dimensional latent space with a projection matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$, resulting in a latent representation $\mathbf{S} \in \mathbb{R}^{k \times N}$. The obtained latent representation is then projected back to the feature space with a projection matrix (embedding function) $\mathbf{W}^* \in \mathbb{R}^{d \times k}$ and becomes $\hat{\mathbf{X}} \in \mathbb{R}^{d \times N}$ i.e. $k < d$ the latent representation/code reduces the dimensionality of the original data input. It is wished that the reconstruction error is minimised, i.e. $\hat{\mathbf{X}}$ is as similar as possible to $\mathbf{X}$. This is achieved by optimising against the following objective:

$$\min_{\mathbf{W},\,\mathbf{W}^*} \|\mathbf{X} - \mathbf{W}^*\mathbf{W}\mathbf{X}\|_F^2 \tag{4.1}$$

4.1.2 Semantic Autoencoder

A conventional autoencoder is unsupervised and the learned latent space has no explicit semantic meaning. With the proposed Semantic AutoEncoder (SAE), it is assumed that each data point also has a semantic representation, e.g., class label or attributes. To make the latent space in the autoencoder semantically meaningful, the simplest approach is taken, that is, the latent space $\mathbf{S}$ is forced to be the semantic space, e.g., each column of $\mathbf{S}$ is now an attribute vector given during training for the corresponding data point. In other words, the latent space is not latent any more during training. The learning objective thus becomes:

$$\min_{\mathbf{W},\,\mathbf{W}^*} \|\mathbf{X} - \mathbf{W}^*\mathbf{W}\mathbf{X}\|_F^2 \quad s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{S} \tag{4.2}$$

To further simplify the model, tied weights are considered as in (Boureau et al., 2008), that is:

$$\mathbf{W}^* = \mathbf{W}^\top$$

The learning objective is then rewritten as follows:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{W}^\top \mathbf{W}\mathbf{X}||_F^2 \quad s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{S} \tag{4.3}$$

Now that only one projection matrix is left to be estimated, instead of two. Furthermore, since $\mathbf{W}\mathbf{X} = \mathbf{S}$, $\mathbf{S}$ is substituted to the first term, and Eq. (4.3) becomes:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{W}^\top \mathbf{S}||_F^2 \quad s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{S} \tag{4.4}$$

Solving an objective with a hard constraint such as $\mathbf{W}\mathbf{X} = \mathbf{S}$ is difficult. Therefore, following standard principle, relaxing the hard constraint into a soft one is considered and the objective is rewritten as:

$$\min_{\mathbf{W}} \underbrace{||\mathbf{W}\mathbf{X} - \mathbf{S}||_F^2}_{\text{Encoder}} + \lambda \underbrace{||\mathbf{X} - \mathbf{W}^\top \mathbf{S}||_F^2}_{\text{Decoder}} \tag{4.5}$$

where λ is a weighting coefficient that controls the importance of the second term which is reconstruction loss. Now it is a good time to reiterate the main idea of the the proposed model in terms of Eq. (4.5): first term corresponds to the encoder which is embedding the data from feature space to the semantic space, while second term is the decoder regenerating the original data from semantic space (see Figure 4.1).

4.1.3 Optimisation

Eq. (4.5) is a convex function hence it has a global optimal solution. By taking simply a derivative of Eq. (4.5) and setting it zero, one can obtain an optimal solution. Specifically, first, Eq. (4.5) is re-organised using trace properties $\text{Tr}(\mathbf{X}) = \text{Tr}(\mathbf{X}^\top)$ and $\text{Tr}(\mathbf{W}^\top \mathbf{S}) = \text{Tr}(\mathbf{S}^\top \mathbf{W})$:

$$\min_{\mathbf{W}} ||\mathbf{W}\mathbf{X} - \mathbf{S}||_F^2 + \lambda ||\mathbf{X}^\top - \mathbf{S}^\top \mathbf{W}||_F^2 \tag{4.6}$$

Then, the derivative of Eq. (4.6) is obtained as follows:

$$\begin{gathered}(\mathbf{W}\mathbf{X} - \mathbf{S})\mathbf{X}^\top - \lambda \mathbf{S}(\mathbf{X}^\top - \mathbf{S}^\top \mathbf{W}) = 0 \\ \lambda \mathbf{S}\mathbf{S}^\top \mathbf{W} + \mathbf{W}\mathbf{X}\mathbf{X}^\top = \mathbf{S}\mathbf{X}^\top + \lambda \mathbf{S}\mathbf{X}^\top \qquad (4.7)\end{gathered}$$

By denoting $\mathbf{A} = \lambda \mathbf{S}\mathbf{S}^\top$, $\mathbf{B} = \mathbf{X}\mathbf{X}^\top$, and $\mathbf{C} = (1 + \lambda)\mathbf{S}\mathbf{X}^\top$, the following equation can be derived:

$$\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{C}, \tag{4.8}$$

which is a well-known Sylvester matrix equation that can be solved efficiently by the Bartels-Stewart algorithm (Bartels and Stewart, 1972) (See for Appendix A for details). In MATLAB, it can be implemented with *a single line* of code: `sylvester`[2]. Importantly, the complexity of

[2] https://uk.mathworks.com/help/matlab/ref/sylvester.html

Algorithm 1 SAE in MATLAB

```
function W = SAE(X,S,lambda)
  % SAE - Semantic AutoEncoder
  % Input:
  %    X: dxN data matrix.
  %    S: kxN semantic matrix.
  %    lambda: regularisation parameter.
  %
  % Return:
  %    W: kxd projection matrix.

  A = lambda*S*S';
  B = X*X';
  C = (1+lambda)*S*X';
  W = sylvester(A,B,C);
end
```

Eq. (4.8) depends on the size of feature dimension ($\mathcal{O}(d^3)$), and not on the number of samples; it thus can scale to large-scale datasets. Algorithm 1 shows a 6-line MATLAB implementation of the proposed solver.

Complexity analysis The complexity of Eq. (4.8) is $\mathcal{O}(d^3)$ as mentioned above. However, before solving Eq. (4.8), we need to obtain $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$. The complexities of $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$ are $\mathcal{O}(k^2N)$, $\mathcal{O}(d^2N)$, and $\mathcal{O}(kNd)$, respectively, where N is the number of training samples, and k is the dimension of semantic space. All in all, the overall complexity is approximately $\mathcal{O}(N)$ in terms of the number of samples if $k \ll N$ and $d \ll N$, while in terms of size of feature dimension it is $\mathcal{O}(d^3)$ if $k \ll d$ and $N \ll d$.

4.1.4 SEA for Cross-class Recognition

Problem formulation Let $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_s\}$ and $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_u\}$ denote a set of s seen and u unseen class labels, and they are disjoint $\mathbf{Y} \cap \mathbf{Z} = \varnothing$. Similarly $\mathbf{S}_Y = \{\mathbf{s}_1, \dots, \mathbf{s}_s\} \in \mathbb{R}^{s \times k}$ and $\mathbf{S}_Z = \{\mathbf{s}_1, \dots, \mathbf{s}_u\} \in \mathbb{R}^{u \times k}$ denote the corresponding seen and unseen class semantic representations (e.g. k-dimensional attribute vector). Given training data with N number of samples

$\mathbf{X}_Y = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i)\} \in \mathbb{R}^{d \times N}$, where $\mathbf{x}_i$ is a d-dimensional visual feature vector extracted from the i-th training image from one of the seen classes, cross-class transfer learning aims to learn a classifier $f : \mathbf{X}_Z \to \mathbf{Z}$ to predict the label of the image coming from unseen classes, where $\mathbf{X}_Z = \{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_i)\}$ is the test data and $\mathbf{z}_i$ and $\mathbf{s}_i$ are unknown.

SAE for cross-class recognition Given semantic representation $\mathbf{S}$ such as attributes, and the training data $\mathbf{X}_\mathbf{Y}$, using the SAE, first the encoder $\mathbf{W}$ and decoder $\mathbf{W}^\top$ are learned by Algorithm 1. Subsequently, cross-class recognition can be performed in two spaces:

1) **Encoder:** With the encoder projection matrix $\mathbf{W}$: a new test sample $\mathbf{x}_i \in \mathbf{X}_Z$ can be embedded to the semantic space by $\hat{\mathbf{s}}_i = \mathbf{W}\mathbf{x}_i$. After that, the classification of the test data in the semantic space can be achieved by simply calculating the distance between the estimated semantic representation $\mathbf{s}_i$ and the embedded prototypes $\mathbf{S}_Z$:

$$\Phi(\mathbf{x}_i) = \arg\min_j D(\hat{\mathbf{s}}_i, \mathbf{S}_{Z_j}) \tag{4.9}$$

where $\mathbf{S}_{Z_j}$ is j-th prototype attribute vector of the j-th unseen class, D is a distance function, and $\Phi(\cdot)$ returns the class label of the sample.

2) **Decoder:** With the decoder projection/embedding matrix $\mathbf{W}^\top$: Similarly, the prototype representations can be embedded to the visual feature space by $\hat{\mathbf{x}}_i = \mathbf{W}^T\mathbf{s}_i$ where $\mathbf{s}_i \in \mathbf{S}_Z$ and $\hat{\mathbf{x}}_i \in \hat{\mathbf{X}}_Z$ is the embedded prototype. Then, the classification of the test data in the feature space can be achieved by calculating the distance between the feature representation $\mathbf{x}_i$ and the prototype embeddings in the feature space $\hat{\mathbf{X}}_Z$:

$$\Phi(\mathbf{x}_i) = \arg\min_j D(\mathbf{x}_i, \hat{\mathbf{X}}_{Z_j}) \tag{4.10}$$

where $\hat{\mathbf{X}}_{Z_j}$ is j-th unseen class prototype embedded in the feature space.

In the experiments it is found that the two testing strategies yield very similar results (see Section 4.2). Results with both strategies are reported unless otherwise specified. Note that LP is not used in this case, since this is an inductive setting.

4.1.5 Relations to Existing Models

Now this thesis reveals interesting relations between the SAE and existing methods. Many existing cross-class transfer learning models learn an embedding function from a visual feature space

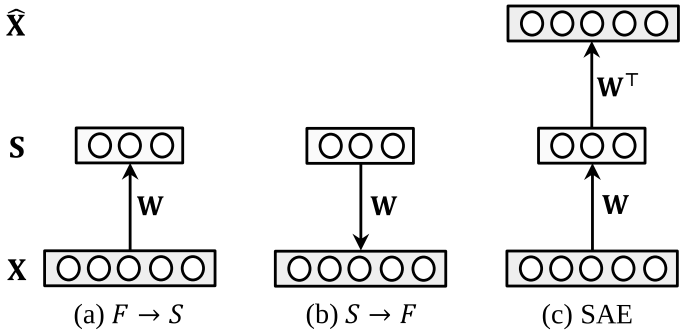


Figure 4.2: Different ways of learning embedding space: (a) F → S, (b) S → F, and (c) Both (SAE). 'F' – Feature space, and 'S' – Semantic space.

to a semantic space (see Figure 4.2(a)). If the embedding function is formulated as a linear ridge regression as follows:

$$\min_{\mathbf{W}} \|\mathbf{W}\mathbf{X} - \mathbf{S}\|_F^2 + \eta \|\mathbf{W}\|_F^2, \tag{4.11}$$

It can be seen that comparing Eq. (4.11) with Eq. (4.5), this is the encoder in the SAE with an additional regularisation term on the embedding matrix **W**.

Recently, Shigeto et al. (2015) proposed to reverse the embedding direction: They embed the semantic prototypes into the features space:

$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}^\top \mathbf{S}\|_F^2 + \eta \|\mathbf{W}\|_F^2 \tag{4.12}$$

so this is the decoder of the SAE but again with the regularisation term to avoid overfitting (see Figure 4.2(b)).

The proposed approach, SAE, can thus be viewed as the combination of both models when ridge regression is chosen as the embedding function and without considering the $\|\mathbf{W}\|_F^2$ regularisation as depicted in Figure 4.2(c). This regularisation is unnecessary in SAE due to the symmetric encoder-decoder design – since $\mathbf{W}^* = \mathbf{W}^\top$, the norm of the encoder embedding matrix $\|\mathbf{W}\|_F^2$ cannot be big because it will then produce large-valued projections in the semantic space, and after being multiplied with the large-norm decoder projection matrix, will result in bad reconstruction. In other words, the regularisation on the norm of the embedding matrices have been automatically taken care of by the reconstruction constraint (Boureau et al., 2008).

4.1.6 Deep SAE

SAE can be extended to a deeper version by adding more layers into SAE (DSAE). Figure 4.3 shows a variant of deep SAE with three mid-level layers. Since introducing more layers means

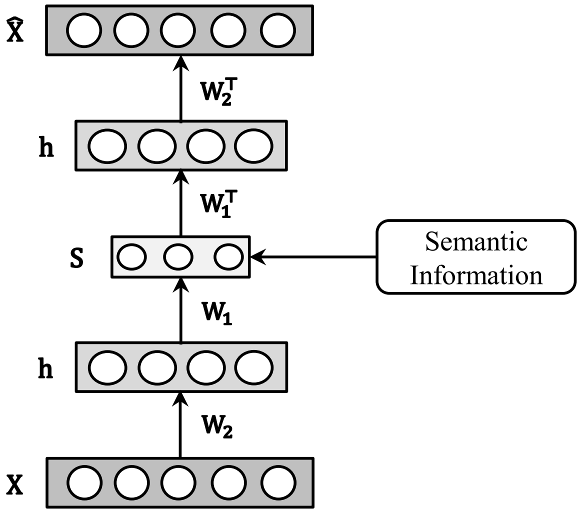


Figure 4.3: Deep semantic autoencoder with three mid-level layers.

learning different weights in different layers, the proposed optimisation method above is inapplicable. Therefore, following the standard deep learning methods, stochastic gradient decent (SGD)[3] is adapted (Goodfellow et al., 2016).

4.2 Experiments and Evaluations

4.2.1 Datasets and Settings

Datasets Six benchmark datasets are used. Four of them are small-scale datasets: AwA (Lampert et al., 2009), CUB (Wah et al., 2011), aPascal&Yahoo (aP&Y) (Farhadi et al., 2009), and SUN (Genevieve et al., 2014). The two large-scale ones are ILSVRC2010 (Deng et al., 2009) (ImageNet-1), and ILSVRC2012/ILSVRC2010 (Russakovsky et al., 2015) (ImageNet-2). In ImageNet-2, as in (Fu and Sigal, 2016), the 1,000 classes of ILSVRC2012 are used as seen classes, while 360 classes of ILSVRC2010, which are not included in ILSVRC2012, for unseen classes. The summary of these datasets is given in Table 2.1.

Semantic spaces Attributes are used as the semantic space for the small-scale datasets, all of which provide the attribute annotations. Semantic word vector representation is used for large-scale datasets. A skip-gram text model is trained on a corpus of 4.6M Wikipedia documents to obtain the word2vec[4] (Mikolov et al., 2013) word vectors.

[3] This can also be applied to optimise SAE objective.

[4] https://code.google.com/p/word2vec/

Small-scale datasets						Large-scale datasets			
Method	SS	AwA	CUB	aP&Y	SUN	Method	SS	ImNet-1	ImNet-2
DAP* (Lampert et al., 2009)	A	60.1	-	38.2	72.0 | 44.5	Rohrbach et al. (2011)	W	34.8	–
ESZSL* (Romera-Paredes and Torr, 2015)	A	75.3	48.7	24.3	82.1 | 18.7	NCM (Thomas et al., 2012)	W	35.7	–
SS-voc (Fu and Sigal, 2016)	A/W	78.3/68.9	-	-	- | –	DeViSE (Frome et al., 2013)	W	31.8	12.8
SJE (Akata et al., 2015)	A+W	73.9	50.1	-	– | 56.1	ConSE (Norouzi et al., 2014)	W	28.5	15.5
MTMDL (Yang and Hospedales, 2015)	A/W	63.7/55.3	32.3	-	- | –	AMP (Fu et al., 2015c)	W	41.0	13.1
SynC^{struct} Changpinyo et al. (2016)	A	72.9	54.4	-	– | 62.7	SS-Voc (Fu and Sigal, 2016)	W	–	16.8
MLZSC (Bucher et al., 2016)	A	77.3	43.3	53.2	84.4 | –				
DS-SJE (Reed et al., 2016a)	A/D	-	50.4/56.8	-	– | –				
AMP (Fu et al., 2015c)	A+W	66.0	-	-	– | –				
DeViSE* (Frome et al., 2013)	A/W	56.7/50.4	33.5	-	– | –				
RRZSL* (Shigeto et al., 2015)	A	80.4	52.4	48.8	84.5 | –				
Lei Ba et al. (2015)	A/W	69.3/58.7	34.0	-	– | –				
SAE (**W**)	A	**84.7**	**61.4**	**55.4**	**91.0** | **65.2**	SAEsee (**W**)	W	**46.1**	**26.3**
SAE ($\mathbf{W}^\top$)	A	**84.0**	**60.9**	**54.8**	**91.5** | **65.2**	SAE ($\mathbf{W}^\top$)	W	**45.4**	**27.2**

Table 4.1: Comparative cross-class recognition accuracy (%, hit@5 for large-scale datasets). For SS (Semantic Space), '/' means 'or' and '+' means 'and'. For CUB, 10 sentence description per image are also used in (Reed et al., 2016a) as input to a language model (word-CNN-RNN) to compute semantic space ('D'). For the SUN dataset, the results are for the 707/10 and 645/72 splits respectively, separated by '|'. '-' means that no reported results are available. **W** parametrises the embedding function of the encoder and $\mathbf{W}^\top$ the decoder. *means that the results are obtained by running the publicly available source codes with GoogLeNet features used in this thesis.

see

Features All recent CCTL methods use visual features extracted by deep convolutional neural networks (DCNNs). In the experiments, GoogleNet (with batch normalisation) is used to extract deep features (Szegedy et al., 2015) which is the 1024D activation of the final pooling layer. The only exception is for ImageNet-1: For fair comparison with published results, Alexnet architecture is used, and it is trained from scratch using the 800 seen classes, resulting in 4096D visual feature vectors computed using the FC7 layer (Krizhevsky et al., 2012).

Implementation details for DSAE TensorFlow toolbox is used with GeForce GTX TITAN X GPU. Adam optimiser is used with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. A learning rate and the number of iterations are set to 0.0002 and 60,000, respectively. All images are scaled to 299×299 pixels.

Parameter settings The SAE model has only one hyperparameter: λ (see Eq. (4.5)). As in (Zhang and Saligrama, 2016), its value is set by class-wise cross-validation using the training data. The dimension of the embedding (middle) layer always equals to that of the semantic space. Only SUN dataset has multiple splits. The same 10 splits are used as in (Changpinyo et al., 2016), and the average performance is reported.

Evaluation metric For the small-scale datasets, multi-way classification accuracy is used as in previous works, while for the large-scale datasets flat hit@K classification accuracy is used as in (Frome et al., 2013). hit@K means that the test image is classified to a 'correct label' if it is among the top K labels. hit@5 accuracy is reported in the experiments as in other works, unless otherwise stated.

Competitors 12 existing cross-class transfer learning models are selected for the small-scale datasets and 6 for the large-scales ones (much fewer existing works reported results on the large-scale datasets). The selection criteria are: (1) Not use of target data: all selected methods do not use unlabelled target data for model learning; (2) recent work: most of them are published in the past several years; (3) competitiveness: they clearly represent the state-of-the-art; and (4) representativeness: they cover a wide range of models.

The brief descriptions of the compared methods are as follows:

- **DAP** (Lampert et al., 2009), **SJE** (Akata et al., 2015), and **AMP** (Fu et al., 2015c): See Section 3.2.2 for details.

- **ESZSL** (Romera-Paredes and Torr, 2015) is an approach that is built on a more general framework which models the relationships between features, attributes, and classes as a two linear layers network, where the weights of the top layer are not learned but are given as a semantic information by the environment. Only attributes are used for cross-class recognition.

- **SynC**^{struct} (Changpinyo et al., 2016) is an approach that is based on manifold learning. The main idea is to align the semantic space deriving from external information to the model space concerning itself with recognizing visual features. For this, a set of 'phantom' object classes are introduced in which their coordinates live in both the semantic space and the model space. The phantom classes serve as a dictionary bases to synthesise real object classifiers. Only attributes are used as semantic information.

- **MLZSC** (Bucher et al., 2016) is a method that attempts to control the semantic embeddings of images using metric learning. To do this, they propose a framework with two constraints: metric discriminating capacity handled by metric learning (Thomas et al., 2012), and accurate attribute prediction. Only attributes are used.

- **DS-SJE** (Reed et al., 2016a) and **Lei Ba et al. (2015)** are methods that are able to use the description that is considered to contain richer information that semantic words. For this, they introduce a model that train end-to-end to align with the fine-grained and category-specific content of images. At higher level, the model consists of mainly two networks: deep convolutional neural network for images and recurrent neural network for description. With two networks, they try to learn a similarity function that align an image with its true description.

- **DeViSE** (Frome et al., 2013) and **MTMDL** (Yang and Hospedales, 2015) are very similar in spirit to DS-SJE, however they use word representation instead of the description, and different network artitechtures.

- **RRZSL** (Shigeto et al., 2015): See Section 2.3 for details.

- **SS-voc** (Fu and Sigal, 2016) is a method based on a maximum margin framework. Specifically, classification is done through nearest neighbor distance to class prototypes in the semantic embedding space, and a set of constraints is encoded ensuring that labelled images project into semantic space such that they become closer to the correct class prototypes than that of incorrect ones.

- **Rohrbach et al. (2011)** DS (Rohrbach et al., 2010) is used for evaluation (see Section 3.2.2 for details).

- **NCM (Thomas et al., 2012)** uses nearest class mean (NCM) as a metric for cross-class recognition.

- **ConSE** (Norouzi et al., 2014) is a simple method for building an embedding system (image) from any existing n-way image classifier and a semantic embedding model (word), which contains the n-class labels in its vocabulary. This method embeds images into the semantic embedding space via convex combination of the class label embedding vectors, and requires no additional training.

As can been known from description of the compared methods, the main difference of SAE compared to comparaed method is that none of them is based on auto-encoder paradigm. Note that the proposed method in the previous chapter (DTCCTL) is not compared with SAE, due to a number of reasons: (1) DTCCTL is transductive; (2) if DTCCTL is considered with no data from the target domain, it is similar to (Shigeto et al., 2015) (RRZSL) in terms of the objective function meaning achieving similar performance with the SAE (provided the same fatures used).

4.2.2 Comparative evaluation

From the results in Table 4.1 the following observations can be made: (1) The SAE achieves the best results on all 6 datasets. (2) On the small-scale datasets, the gap between the SAE's results to the strongest competitor ranges from 3.5% to 6.5%. This is despite the fact that most of the compared models use far complicated nonlinear models and some of them use more than one semantic space. (3) On the large-scale datasets, the gaps are even bigger: On the largest ImageNet-2, the proposed model improves over the state-of-the-art SS-Voc by 8.8%. (4) Both the encoder and decoder embedding functions in the SAE ($\mathbf{W}$ and $\mathbf{W}^{\top}$) respectively) can be used for effective cross-class recognition. The encoder projection function seems to be slightly better overall.

4.2.3 Ablation Study

The key strength of the SAE comes from the additional reconstruction constraint in the autoencoder formulation. Since most existing cross-class transfer learning models use more sophisticated embedding/projection functions than the proposed linear mapping, in order to evaluate how important this additional constraint is, CCTL baselines that use the same simple embedding functions as the SAE are considered. As discussed in Section 4.1.5, without the constraint both the encoder and decoder can be considered as conventional CCTL models with linear ridge regression as embedding function, and they differ only in the embedding directions. Table 4.2 shows that, when the embedding function is the same, adding the additional reconstruction constraint makes a huge difference. Note that comparing to the state-of-the-art results in Table 4.1, simple ridge regression is competitive but clearly inferior to the best models due to its simple linear embedding function. However, when the two models are combined in the SAE, a much more powerful model is obtained in which it beats all existing models.

Projection	AwA	CUB	aP&Y	SUN
F → S	60.6	41.1	30.5	71.5
F ← S	80.4	52.4	48.8	84.5
SAE	**84.7**	**61.4**	**55.4**	**91.0**

Table 4.2: The importance of adding the reconstruction constraint. Both compared methods are based on ridge regression and differ in the embedding direction between the visual and semantic spaces. Attributes are used. The encoder is used (see Section 4.1.4).

4.2.4 DSAE results

For DSAE, experiments on AwA and CUB datasets are conducted. The number of mid-level layers are varied from 3 to 7. Table 4.3 shows the results, and the following observations are made: (1) DSAE can learn better encoder function with enough amount of data. For example, DSAE with three mid-level layers (DASE-3) outperforms SAE with about 1.5% increase on AwA. However, with less amount of data this performance gain cannot be accomplished which is the case on CUB dataset, with slight .1% increase; (2) However, as expected the time for training increases considerably, as the number of layers increases: {90, 140, and 167} seconds for DSAE-

{3, 5, and 7}, respectively while SAE takes only 1.3 seconds. (3) By introducing more layers, the model performance decreases. That is, the network architecture becomes more complex, and overfits the training data. (4) It is noted that there are cases that mini-batch stochastic gradient descent gives slightly higher performance than the proposed optimisation in AwA dataset: 84.7 % (this thesis), and 85.3 (SGD). Indeed, there are several things which could be applied to improve the generalisation of the network such as batch normalisation, different activation functions, and dropout regularisation. Analysis of different techniques are left for future work.

AwA				
	SAE	DSAE-3	DSAE-5	DSAE-7
Accuracy (%)	85.3	**87.0**	86.4	85.0
Training Time (s)	10.3	90	140	167
CUB				
	SAE	DSAE-3	DSAE-5	DSAE-7
Accuracy (%)	61.4	**61.5**	58.5	58.6
Training Time (s)	5.2	66	104	131

Table 4.3: Deep SAE results on AwA and CUB. Accuracy and training time are reported. DSAE-3 - ’3’ refers to the number of mid-level layers.

4.2.5 Further Analysis

Generalised cross-class transfer learning Another cross-class transfer learning setting that emerges recently is the generalised setting[5] under which the test set contains data samples from both the seen and unseen classes. By following the same setting of (Chao et al., 2016b), 20% of the data samples from the seen classes are held out and they are mixed with the data samples from the unseen classes. In this setting, the evaluation metric is Area Under Seen-Unseen accuracy Curve (AUSUC), which measures how well a cross-class recognition method can trade-off between recognising data from seen classes and that of unseen classes (Chao et al., 2016b). The upper bound of this metric is 1, not the same as the accuracy (%) used for cross-class classification. The results on AwA and CUB are presented in Table 4.4 comparing the SAE with 5 other alternatives. It can be seen that on AwA, the SAE is slightly worse than the state-of-the-art method SynC^{struct}. However, on the more challenging CUB dataset, the proposed method

[5] This setting is also called generalised zero-shot learning/recognition in the literature.

Method	AwA	CUB
DAP (Lampert et al., 2009)	0.366	0.194
IAP (Lampert et al., 2009)	0.394	0.199
ConSE (Norouzi et al., 2014)	0.428	0.212
ESZSL (Romera-Paredes and Torr, 2015)	0.449	0.243
SynC^{struct} (Changpinyo et al., 2016)	**0.583**	0.356
SAE	0.579	**0.448**

Table 4.4: Comparative evaluation on generalised cross-class transfer learning on AwA and CUB. Encoder is used. Note that Area Under Seen-Unseen accuracy Curve (AUSUC) metric is used, and upper bound for this metric is 1.

significantly outperforms the competitors.

Computational cost The following experiments were conducted in MATLAB on a PC with two 3.40 GHz CPUs and 16G RAM. The computational cost of SAE is evaluated with comparison to two linear CCTL models: ESZSL and AMP which are among the more efficient existing CCTL models (all of them are implemented in MATLAB). Table 4.5 shows that for model training, the SAE is at least 10 times faster. For testing, the SAE is still the fastest, although ESZSL is close.

Method	Training (in seconds)	Test (in seconds)
ESZSL (Romera-Paredes and Torr, 2015)	16	0.08
AMP (Fu et al., 2015c)	844	0.23
SAE	1.3	0.07

Table 4.5: Evaluating the computational cost (in seconds) on AwA. Encoder is used.

4.3 Conclusion

In this chapter, a novel cross-class transfer learning model based on a semantic autoencoder (SAE) is proposed. The SAE is built on an encoder-decoder paradigm. It is very simple and computationally fast linear embedding function, and introduces an additional reconstruction objective function for learning a more generalisable embedding function. Very efficient optimisation for the proposed objective function is also proposed in which the solution of the objective function requires only solving Sylvester matrix equation. Importantly, the complexity of the solution is analysed and it is shown that it is only dependent on the feature dimension, not the number of samples. The interesting connection to existing methods are demonstrated as well e.g., ridge regression, reverse regression. It is demonstrated through extensive experiments that this new SAE model outperforms existing standard cross-class transfer learning methods and generalised cross-class transfer learning methods on four mid-scale and two large-scale benchmark datasets in terms of recognition accuracy and computational cost. Also, the SAE is extended to deep version named deep SAE by introducing hidden layers into the SAE architecture. Early experimental results indicate that the deep SAE could perform better in some datasets, because it could learn more complex an embedding function.

Chapter 5

Unsupervised Cross-class Transfer Learning by L_1 Graph Learning

The preceding Chapters (3 and 4) describe transductive and inductive cross-class transfer learning methods. They both fall into supervised learning paradigm, meaning they require labelled data for model learning. Because labels can be used to connect seen and unseen classes in semantic space, those methods are suitable for category recognition. In contrast, this chapter deals with the case that no labelled data is available during model learning, meaning there is no way of connecting seen class and unseen classes. Therefore, the category recognition cannot be performed under this setting. However, this setting can be used for instance recognition, whose goal is a matching image pairs. The matching largely requires robust and discriminative representation so that it can be successfully accomplished. To this end, this chapter proposes a novel method for the unsupervised cross-class transfer learning with application to person re-identification (ReID).

Most recent ReID methods are based on supervised learning Wei et al. (2017); Xiao et al. (2016). Given a set of labelled training data consisting of images of people paired across camera views according to identity, a distance metric is learned either using hand-crafted features or end-to-end fashion using deep learning methods Zheng et al. (2016); Xiong et al. (2014a); Ahmed et al. (2015). However, they require images of hundreds or more people to be paired across each pair of camera views which is both tedious and sometimes not possible – some people do not reappear in other camera views. This severely limits the scalability of the existing methods making them unsuitable for practical large scale ReID tasks. To overcome this problem, a number

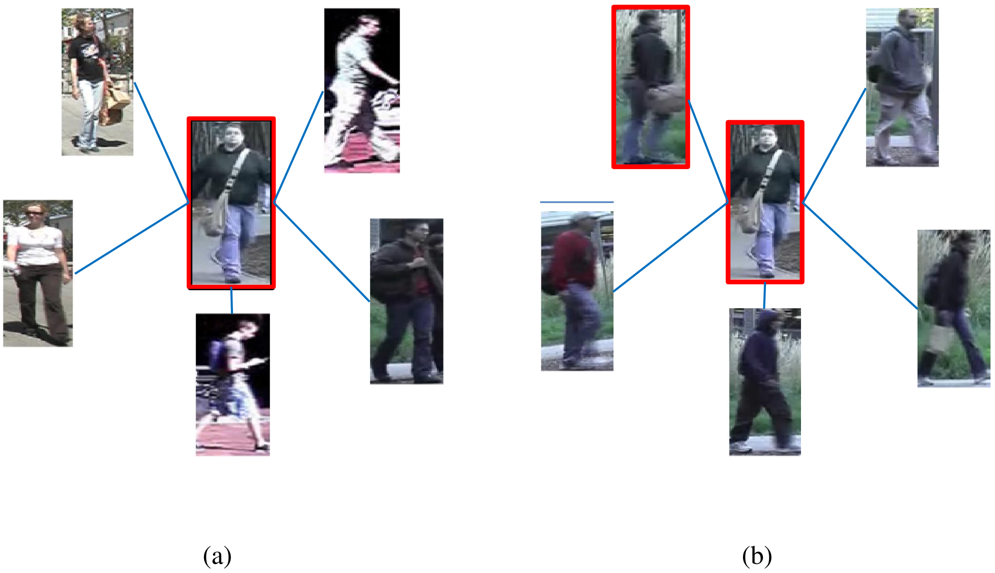


Figure 5.1: An illustration of graph learning for ReID. (a) A graph constructed in the hand-crafted feature space; (b) A graph learned using the proposed model in this work. One graph node and its five connected neighbours are shown, with the neighbour capturing the same person highlighted in red. See Figure 5.5 for the actual graph encoded in a similarity matrix $\mathbf{W}$.

of unsupervised ReID methods have been proposed Zhao et al. (2013a); Wang et al. (2014); Hehe et al. (2017); Wang et al. (2016a). However, without labelled training data, they can only focus on learning salient and view invariant representations. Their performance is thus much weaker compared to the supervised methods. This is because they are unable to learn the cross-view discriminative information effectively, critical for matching the same person whilst separating the person from imposters of similar appearance. Due to their uncompetitiveness in published benchmarking metrics, these unsupervised learning models have received little attention when practicality and scalability are not considered in current benchmarking. It is true that recently more and more large labelled datasets are becoming available publicly Zheng et al. (2015a). However, they could not be used at this point of time, due to a bias of the dataset. That is, every dataset is collected under a certain environment, thus learning a model with those datasets and applying the model to a new environment does not generalise well Peng et al. (2016).

This chapter introduces a novel model that can learn discriminative low-dimensional feature representation from a set of unlabelled data that can cope well with view-invariance. The proposed method is built on dictionary learning models that are shared across camera views. It is easy to understand how a representation obtained by dictionary learning can be view-invariant and low-dimensional – dictionary learning is widely used as an unsupervised model for dimensionality reduction (Kenneth et al., 2003; Aharon et al., 2006); and by sharing the same dictionary across camera views, it intrinsically requires that the learned representation to be view-invariant. It is the discriminative part that is non-trivial: How can we enforce that the learned representation is good for matching people across camera views, without the discriminative information from a set of paired training data?

The solution proposed in this chapter is to relax the definition of discriminativity. Consider each dictionary atom as a new feature dimension, a learned dictionary defines a subspace, into which the original data points represented by high-dimensional hand-crafted feature vectors are projected. Instead of enforcing that data points corresponding to the *same* person to be as close as possible whilst being further away from other people in the learned subspace as in supervised learning, the visually *similar* people are constrained to be close to each other. Without identity labels, this is obviously a weaker constraint but the best available. Specifically, discriminativity is achieved unsupervised via a visual similarity constraint, which is enforced by introducing a graph Laplacian regularisation term in the dictionary learning objective function (Nie et al., 2011).

However, two problems remain when the conventional graph Laplacian constraint is used in the ReID: (1) The conventional term has a squared l_2-norm, which makes the term susceptible to data outliers. This is particularly unsuitable for the ReID problem as there are plenty of data outliers in ReID, caused by various reasons such as the person detection boxes being imperfect and severe (self-)occlusions. (2) The visual similarity is encoded in a graph whose topology and edge weights are all determined by distances computed using the original high-dimensional hand-crafted features. However, these features are not ideal for people matching, hence learning a new representation in the first place. As illustrated in Figure 5.1(a), a graph constructed using the hand-crafted features connects many visually dissimilar neighbours to each node. This diminishes the power of the graph regularisation term as a visual similarity constraint.

To overcome these two problems, this chapter introduces robust graph regularisation term, and propose to learn the new representation and the optimal graph jointly. Specifically, a l_1-norm is introduced in the proposed graph regularisation term to make it robust against outliers. With this l_1-norm and joint graph and dictionary learning, the learning objective function is both non-smooth and non-convex. Solving this optimisation problem is thus non-trivial. An efficient iterative optimisation algorithm is formulated in this work to solve it. Once learned, the proposed model can be used to compute a representation for each image much more efficiently than any existing unsupervised ReID method. The final matching is done by computing a simple cosine distance between a pair of the representation vectors. Moreover, the proposed method requires very weak data annotation: since cross-view information is crucial for ReID, graph is constructed by restricting the graph edges to connecting cross-view nodes only.

The remainder of this chapter is structured as follows. Section 5.1 provides the details of the proposed graph regularised dictionary learning framework for ReID. Comprehensive experiments and evaluations with comparison to contemporary ReID methods are presented in Section 5.2. Finally, concluding remarks are presented in Section 5.3.

5.1 Methodology

5.1.1 Problem Formulation

Suppose a set of *unlabelled* training examples collected from two camera views[1]. They are denoted as $\mathbf{X} = [\mathbf{X}^a,\ \mathbf{X}^b] \in \mathbb{R}^{n \times m}$, where $\mathbf{X}^a = [\mathbf{x}^a_1,\ \dots,\mathbf{x}^a_{m_1}] \in \mathbb{R}^{n \times m_1}$ contains n-dimensional feature vectors of m_1 images in view A, and $\mathbf{X}^b = [\mathbf{x}^b_1,\ \dots,\mathbf{x}^b_{m_2}] \in \mathbb{R}^{n \times m_2}$ of m_2 images in view B, thus having $m = m_1 + m_2$ data examples in total. The objective of unsupervised person ReID is to learn a matching function f from $\mathbf{X}$, so that given $\mathbf{x}^a$ and $\mathbf{x}^b$ as two test person images from A and B respectively, $f(\mathbf{x}^a, \mathbf{x}^b)$ can match their identities.

5.1.2 Robust Graph Regularised Dictionary Learning

The problem defined above is solved by learning a dictionary $\mathbf{D} \in \mathbb{R}^{n \times k}$ shared by the two camera views using $\mathbf{X}$. Every atom of the learned dictionary (column of $\mathbf{D}$) can be considered as a latent appearance attribute that is invariant to camera view condition changes. Therefore, with this dictionary, each n-dimensional hand-crafted feature vector, regardless which view it comes from, is represented by the coefficients of the k dictionary atoms. This is equivalent to projecting the original n-dimensional hand-crafted feature vectors to a lower-dimensional ($k < n$) latent attribute space. The matching is done by computing a simple cosine distance between two coefficient vectors in the space. Formally, the aim is to learn the optimal dictionary $\mathbf{D}$, such that the latent attribute representation of $\mathbf{X}$, denoted as $\mathbf{S} = [\mathbf{S}^a,\ \mathbf{S}^b] \in \mathbb{R}^{k \times m}$, where $\mathbf{S}^a = [\mathbf{s}^a_1,\ \dots,\mathbf{s}^a_{m_1}] \in \mathbb{R}^{k \times m_1}$ and $\mathbf{S}^b = [\mathbf{s}^b_1,\ \dots,\mathbf{s}^b_{m_2}] \in \mathbb{R}^{k \times m_2}$, are optimised for matching the training data. It is expected that the same $\mathbf{D}$ can be generalised to match unseen test data across camera views.

[1] In practice the proposed model is not restricted by the number of camera views. Two here is used purely for notational simplicity.

Conventional dictionary learning methods estimate the dictionary $\mathbf{D}$ and the representation $\mathbf{S}$ simultaneously by solving the following optimisation problem:

$$(\mathbf{D}^*, \mathbf{S}^*) = \min_{\mathbf{D},\mathbf{S}} \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda_1 \Omega(\mathbf{S}) \;\; s.t. \;\; \|\mathbf{d}_i\|_2^2 \leq 1, \tag{5.1}$$

where $\|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2$ is the reconstruction error evaluating how well a linear combination of the learned atoms can approximate the input data, and $\|\cdot\|_F$ denotes the matrix Frobenious norm. $\Omega(\mathbf{S})$ is a regularisation term that is weighted by λ_1. Existing models differ mainly in the choice of the regularisation term on $\mathbf{S}$. The sparsity term, $\Omega(\mathbf{S}) = \|\mathbf{S}\|_1$ is widely used which favours a small number of atoms for reconstruction. The constraint $\|\mathbf{d}_i\|^2 \leq 1$ ($\mathbf{d}_i$ is a column of $\mathbf{D}$, $i = 1, ..., k$) enforces the learned dictionary atoms to be compact. It is clear from this formulation that a conventional dictionary learning model only cares about how to best reconstruct $\mathbf{X}$ using $\mathbf{D}$ and $\mathbf{S}$, without taking into account whether the representation $\mathbf{S}$ is discriminative. For learning a discriminative dictionary for cross-view ReID, one must exploit cross-view identity discriminative information.

A learned dictionary can be made discriminative by using a graph regularisation term which dictates that visually similar people will be close to each other in the learned latent attribute space (Chung, 1997). Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be an undirected graph connecting between the data points where $\mathbf{V}$ and $\mathbf{E}$ are a set of graph vertices representing the data points and an edge set, respectively. This graph can be encoded by an affinity matrix $\mathbf{W} \in \mathbb{R}^{m \times m}$ for m data points where $\mathbf{W}_{ij} \neq 0$ if the two vertices are connected, *i.e.* the corresponding data points are in a local neighbourhood. Note: (1) In the context of ReID, the goal is to learn the cross-view discriminative dictionary, thus restricting the graph edges to connecting cross-view nodes only. (2) the graph regularisation term is used to replace the commonly used sparsity constraint $\|\mathbf{S}\|_1$, for reasons to be explained later (see robust graph regularisation paragraph). A standared graph regularisation term $\Omega(\mathbf{S})$ is defined as:

$$\Omega(\mathbf{S}) = \sum_{ij}^{m} \mathbf{W}_{ij} \|\mathbf{s}_i - \mathbf{s}_j\|_2^2. \tag{5.2}$$

This regularisation essentially requires that the projected data points in the learned latent attribute space to be smooth with regards to the graph, that is, their distances need to conform to the visual similarity relationship embedded in the graph. However, it is found that Eq. (5.2) has two critical limitations that make it unsuitable for the unsupervised ReID problem. First, the distance between two projected data points is calculated with a *squared l_2-norm*. It is well-known that a square-based regularisation function can be easily dominated by outlying data samples. Unfortunately outlying samples are commonplace in ReID because of background in person detection bounding boxes, detector errors, and (self-)occlusions. Another limitation arises from how the graph is constructed. Most existing methods build the graph in the original high dimensional hand-crafted feature space using $\mathbf{X}$. This is suboptimal – if the hand-crafted feature space is good for measuring cross-camera visual similarity, the ReID problem would have already been solved. Learning a discriminative latent attribute space is precisely due to the fact that measuring visual similarity in the original space is unreliable and error-prone, as illustrated in Figure 5.1. To tackle both limitations simultaneously, this chapter introduces a robust graph regularisation formulation and a joint graph and dictionary learning method.

Robust graph regularisation. This new term is designed to alleviate the effect of outlying samples during model learning. To derive the robust graph regularisation, let's first rewrite Eq. (5.2) in a matrix form with trace notation:

$$\Omega(\mathbf{S}) = \sum_{ij}^{m} \mathbf{W}_{ij} \|\mathbf{s}_i - \mathbf{s}_j\|_2^2 = \mathrm{Tr}(\mathbf{S}\mathbf{L}_{\mathbf{W}}\mathbf{S}^\top) \tag{5.3}$$

where $\mathbf{L}_{\mathbf{W}} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix, $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ is a degree matrix. Let $\mathbf{L}_{\mathbf{W}} = \mathbf{U}_{\mathbf{W}}\mathbf{H}_{\mathbf{W}}\mathbf{U}_{\mathbf{W}}^\top$ using the eigen-decomposition technique, and after some matrix manipulation:

$$\mathrm{Tr}(\mathbf{S}\mathbf{L}_{\mathbf{W}}\mathbf{S}^\top) = \mathrm{Tr}(\mathbf{S}\mathbf{U}_{\mathbf{W}}\mathbf{H}_{\mathbf{W}}\mathbf{U}_{\mathbf{W}}^\top\mathbf{S}^\top) = \mathrm{Tr}(\mathbf{S}\mathbf{U}_{\mathbf{W}}\mathbf{H}_{\mathbf{W}}^{\frac{1}{2}}\mathbf{H}_{\mathbf{W}}^{\frac{1}{2}}\mathbf{U}_{\mathbf{W}}^\top\mathbf{S}^\top) = \|\mathbf{S}\mathbf{A}_{\mathbf{W}}\|_F^2 \tag{5.4}$$

where $\mathbf{A}_{\mathbf{W}} = \mathbf{U}_{\mathbf{W}}\mathbf{H}_{\mathbf{W}}^{\frac{1}{2}}$. Eq. (5.4) above is quadratic. To promote sparsity and suppress effects of outlying samples, a l_1-norm is adopted instead of the Frobenius norm James et al. (2013); Kim et al. (2009). This gives the proposed graph weighted l_1-norm regularisation term:

$$\Omega_{\mathrm{R1}}(\mathbf{S}) = \|\mathbf{S}\mathbf{A}_{\mathbf{W}}\|_1. \tag{5.5}$$

Replacing $\Omega(\mathbf{S})$ with $\Omega_{\mathrm{R1}}(\mathbf{S})$ in Eq. (5.1), a robust graph regularised dictionary learning model is obtained:

$$\min_{\mathbf{D},\mathbf{S}} \quad \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda_1 \|\mathbf{S}\mathbf{A}_{\mathbf{W}}\|_1 \;\; s.t. \;\; \|\mathbf{d}_i\|^2 \leq 1 \tag{5.6}$$

The key advantages of the proposed robust graph regularisation in this work over the conventional regularisation formulation, including the existing dictionary learning based ReID model DLLAP (Kodirov et al., 2015), and UCDTL (Peng et al., 2016), are as follows:

1. *Non-linearity.* Robust graph regularisation introduces non-linearity into the objective, i.e. $\mathbf{S}$ is non-linear with respect to the original data matrix $\mathbf{X}$, whilst the conventional graph regularisation is linear.

2. *Sparsity.* It is well-known that l_1-norm has a shrinkage property thus promotes sparsity (James et al., 2013; Kim et al., 2009). Intuitively, in the presence of noise and outliers, the magnitude of $||\mathbf{SA}_W||_F^2$ of the regularisation becomes very big for those outlying data points, and as a result the whole objective function could be dominated by the noise and outliers. In contrast, $||\mathbf{SA}_W||_1$ becomes sparse due to the use of l_1-norm, consequently suppressing the impact of outliers and noises. Moreover, in the proposed robust regularisation model, explicit sparsity constraint such as $||\mathbf{S}||_1$ is no longer needed[2].

Joint graph and dictionary learning. Instead of computing $\mathbf{W}$ using $\mathbf{X}$ and fixing it during model learning, it is assumed that $\mathbf{W}$ (hence the graph $\mathbf{G}$ as $\mathbf{W}$ depends on the topology of $\mathbf{G}$) is unknown and has to be learned together with $\mathbf{D}$ and $\mathbf{S}$. The objective function thus becomes:

$$\begin{aligned}&\min_{\mathbf{D},\mathbf{W},\mathbf{S}} \frac{1}{2}||\mathbf{X}-\mathbf{DS}||_F^2 + \lambda_1||\mathbf{SA}_\mathbf{W}||_1 + \lambda_2||\mathbf{W}||_F^2 \\ &s.t. \quad ||\mathbf{d}_i||_2^2 \leq 1,\ \mathbf{W}_i^\top \mathbf{1} = 1,\ \mathbf{W}_i \geq 0.\end{aligned} \tag{5.7}$$

where $\lambda_2||\mathbf{W}||_F^2$ is a regularisation term on $\mathbf{W}$ weighted by λ_2 to prevent trivial solutions. The constraints, $\mathbf{W}^\top\mathbf{1} = 1$ and $\mathbf{W} \geq 0$, ensure the validity of the learned graph. It is shown in the experiments (Section 5.2.2) that this novel joint learning of graph and dictionary has significant advantage over the existing dictionary learning based ReID model DLLAP (Kodirov et al., 2015). In the following, the proposed model is referred to as L_1Graph for convenience unless otherwise stated.

5.1.3 Optimisation

The optimisation problem in Eq. (5.14) is non-convex and non-smooth. Solving it is thus more difficult than Eq. (5.1) due to the l_1-norm used in $\Omega_{\text{R1}}(\mathbf{S})$ and the additional unknown variable $\mathbf{W}$. Next, an efficient solver is developed for Eq. (5.14) based on the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011).

[2] Empirically it is found in this work that adding an extra $||\mathbf{S}||_1$ term makes little difference to the ReID performance, but results in more complex solver and higher computational cost.

First, Eq. (5.14) is transformed by letting $\mathbf{U} = \mathbf{S}\mathbf{A}_{\mathbf{W}}$, then the Augmented Lagrangian function of Eq. (5.14) with the introduced constraint is:

$$\begin{aligned} \mathcal{L}_{(\mathbf{D},\mathbf{S},\mathbf{U},\mathbf{W})} =& \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{S}\|_F^2 + \lambda_1\|\mathbf{U}\|_1 + \langle \mathbf{F}, \mathbf{U}-\mathbf{S}\mathbf{A}_{\mathbf{W}}\rangle \\ &+ \frac{\gamma}{2}\|\mathbf{U}-\mathbf{S}\mathbf{A}_{\mathbf{W}}\|_F^2 + \lambda_2\|\mathbf{W}\|_F^2 \\ s.t. \qquad & \|\mathbf{d}_i\|^2 \leq 1,\ \mathbf{W}^\top\mathbf{1} = 1,\ \mathbf{W} \geq 0. \end{aligned} \tag{5.8}$$

where $\mathbf{F}$ is Lagrangian multiplier, and γ is a penalty parameter. Now, it can be solved alternatingly with the following five steps with respect to $\mathbf{D}$, $\mathbf{S}$, $\mathbf{U}$, and $\mathbf{W}$, respectively. $\mathbf{S}$, $\mathbf{F}$, and $\mathbf{U}$ are initialised randomly in the first iteration.

1) Solving for D: To learn $\mathbf{D}$ for a given $\mathbf{S}$, the objective function reduces to:

$$\min_{\mathbf{D}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{S}\|_F^2 \quad s.t.\ \|\mathbf{d}_i\|_2^2 \leq 1 \tag{5.9}$$

To solve this, the Lagrange dual method is used as in (Lee et al., 2006). The analytical solution of $\mathbf{D}$ can be computed as: $\mathbf{D}^* = \mathbf{X}\mathbf{S}^{\mathbf{T}}(\mathbf{S}\mathbf{S}^\top + \Lambda)^{-1}$, where Λ is a diagonal matrix constructed from all the optimal dual variables.

2) Solving for S: For given $\mathbf{D}$, $\mathbf{F}$, $\mathbf{W}$, and $\mathbf{U}$, solve the following objective to estimate $\mathbf{S}$:

$$\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{S}\|_F^2 + \frac{\gamma}{2}\|\mathbf{U}-(\mathbf{S}\mathbf{A}_{\mathbf{W}} - \frac{\mathbf{F}}{\gamma})\|_F^2.$$

Since each term in this objective is quadratic, its derivative can be obtained and it is set to zero which gives

$$(\mathbf{D}^{\mathbf{T}}\mathbf{D}\mathbf{S} + \gamma\mathbf{S}\mathbf{A}_{\mathbf{W}}\mathbf{A}_{\mathbf{W}}^\top) = \mathbf{D}^\top\mathbf{X} + \gamma\mathbf{U}\mathbf{A}_{\mathbf{W}}^\top + \mathbf{F}\mathbf{A}_{\mathbf{W}}^\top.$$

This is a standard Sylvester equation, which is solved using the Bartels-Stewart algorithm (Bartels and Stewart, 1972).

3) Solving for U: For a given $\mathbf{S}$, solve the following objective to estimate $\mathbf{U}$:

$$\min_{\mathbf{U}} \lambda_1\|\mathbf{U}\|_1 + \frac{\gamma}{2}\|\mathbf{U}-(\mathbf{S}\mathbf{A}_{\mathbf{W}} - \frac{\mathbf{F}}{\gamma})\|_F^2.$$

The soft-thresholding operator can be used to get **U**:

$$\mathbf{U} = \text{sign}\left(\mathbf{SA_W} - \frac{\mathbf{F}}{\gamma}\right)\max(\left|\mathbf{SA_W} - \frac{\mathbf{F}}{\gamma}\right| - \frac{\lambda_1}{\gamma}). \tag{5.10}$$

4) Solving for W: Given **S**, the objective function with respect to **W** is:

$$\min_{\mathbf{W}} \lambda_1 \sum_{ij}^{m} \mathbf{W}_{ij}\|\mathbf{s}_i - \mathbf{s}_j\|_1 + \lambda_2\|\mathbf{W}\|_F^2 \;\; s.t. \;\; \mathbf{W}_i^\top \mathbf{1} = 1, \mathbf{W}_i \geq 0.$$

$\lambda_1 = 1$ is chosen for easiness, and denote $\mathbf{d}_{ij} = \frac{\|\mathbf{s}_i - \mathbf{s}_j\|_1}{2\lambda_2}$ and $\|\mathbf{W}\|_F^2 = \sum_{ij}\mathbf{W}_{ij}^2$, then

$$\min_{\mathbf{W}} \sum_{ij}^{m} \mathbf{W}_{ij}\mathbf{d}_{ij} + \sum_{ij}^{m} \mathbf{W}_{ij}^2 \;\; s.t. \;\; \mathbf{W}_i^\top \mathbf{1} = 1, \mathbf{W}_i \geq 0.$$

The above optimisation problem is composed of independent problems with respect to i, and therefore can be rewritten in a vector form:

$$\min_{\mathbf{W}_i} \|\mathbf{W}_i + \mathbf{d}_i\|_2^2 \;\; s.t. \;\; \mathbf{W}_i \mathbf{1} = 1, \mathbf{W}_i \geq 0.$$

There is a closed-form

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{W},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1\|\mathbf{SA_W}\|_1 + \lambda_2\|\mathbf{W}\|_F^2 \\ &s.t. \;\; \|\mathbf{d}_i\|_2^2 \leq 1, \; \mathbf{W}_i^\top \mathbf{1} = 1, \; \mathbf{W}_i \geq 0. \end{aligned} \tag{5.11}$$

solution using Lagrange multipliers (Nie et al., 2014) for this problem:

$$\mathbf{W}_i = \left(\frac{1 + \sum_{j=1}^{K} \tilde{\mathbf{d}}_j}{K}\mathbf{1} - \mathbf{d}_i\right)_+ \tag{5.12}$$

where the operator $(\mathbf{q})_+$ projects negative elements in $\mathbf{q}$ to 0. K is the parameter that controls the number of neighbours. $\tilde{\mathbf{d}}_i$ is $\mathbf{d}_i$ but with ascending order. After obtaining **W**, it is symmetrised, and eigen-decomposition is done to get $\mathbf{U_W}$ and $\mathbf{S_W}$. Then, $\mathbf{A_W} = \mathbf{U_W}\mathbf{H}_{\mathbf{W}}^{\frac{1}{2}}$. Note that the regularisation parameter λ_2 can be determined by (Nie et al., 2014):

$$\lambda_2 = \frac{1}{m}\sum_{i=1}^{m}(\frac{K}{2}\mathbf{d}_{i,K+1} - \frac{1}{2}\sum_{j=1}^{K}\mathbf{d}_{ij}). \tag{5.13}$$

5) Updating multipliers: F, γ,

$$\mathbf{F} = \mathbf{F}^{old} + \gamma(\mathbf{U} - \mathbf{S}\mathbf{A}_{\mathbf{W}}), \;\; \gamma = \rho\gamma^{old}$$

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{W},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda_1\|\mathbf{S}\mathbf{A}_{\mathbf{W}}\|_1 + \lambda_2\|main_formulation_final\mathbf{W}\|_F^2 \\ &s.t. \;\; \|\mathbf{d}_i\|_2^2 \leq 1, \; \mathbf{W}_i^\top\mathbf{1} = 1, \; \mathbf{W}_i \geq 0. \end{aligned} \tag{5.14}$$

In this work, ρ is set to 1.1, while 0.1 for γ. Typically the value for ρ is set between 1.0 and 1.8 (Boyd et al., 2011).

5-step iteration is continued for **D**, **S**, **U**, **W** until a maximum number of iterations is reached or a predefined threshold (10^{-3}) is satisfied.

Convergence Analysis. The theoretical convergence proof of ADMM does not exist. However, in practice it is guaranteed that the objective function converges to at least a stable point (Boyd et al., 2011). This is validated by the experiments. In particular, it is observed that the proposed algorithm has a stable convergence behaviour, in all tested datasets converging after 10-25 iterations (see Figure 5.3).

Remark on Computational Complexity and Scalability. The optimisation problem of the proposed model (Eq. (5.8)) is solved by dividing it into five optimisation subproblems and solving them alternatingly. Among those subproblems, there are three operations that give rise to high computational cost, whilst the remaining parts have linear complexity $O(m)$, where m is the number of samples. The three operations are: (1) Before solving any of the 5 subproblems, Eigendecomposition of the Laplacian matrix needs to be performed, **L**, in order to derive Eq. (5.4) from Eq. (5.3), which has a complexity of $O(m^3)$; (2) The second subproblem ("Solving for **S**") requires solving of Sylvester equation which has a complexity of $O(m^3)$; (3) Note that before solving the Sylvester equation matrix multiplication operation of $\mathbf{A_w}\mathbf{A_w^T}$ needs to be computed. This operation also has a complexity of $O(m^3)$.

Since the aforementioned operations are very common in the matrix analysis and numerical algebra fields, fast solutions have been considered in several previous works. Specifically, for the first one, since the Laplacian matrix is sparse, specific sparse analysis methods can be employed as in (Fokkema et al., 1998). In this way, the complexity can be reduced to $O(rm^2)$, where r is the ratio of nonzero samples in **L** to the total number of samples m and in this work this ratio is usually very small e.g., it is 0.016 for VIPeR dataset. For the second one which is solving a Sylvester equation, the recently proposed method can be readily adopted in (Rao et al., 2015) which reduces a complexity of $O(m^2)$. For the last one, the eigendecomposition of the $\mathbf{L}$ matrix can be leveraged. That is, the first k largest eigenvalues are chosen to get $\mathbf{A_W}$, and k can be very small compared to the number of samples – $\mathbf{A_W} \in \mathbb{R}^{m\times m}$ matrix becomes $\mathbf{A_W} \in \mathbb{R}^{m\times k}$. With the reduced $\mathbf{A_W}$, the cost is $O(km^2)$. All in all, the fact that all three operations have a complexity of $O(m^2)$ brings down the complexity of the whole algorithm to $O(m^2)$.

5.1.4 Bayesian Interpretation

The conventional sparse coding formulation is:

$$\min_{\mathbf{D},\mathbf{S}} \frac{1}{2} \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda \|\mathbf{S}\|_1 \tag{5.15}$$

It is straightforward that Eq. (5.15) has an equivalent interpretation in the Bayesian framework according to Huang and Aviyente (2007); Bishop et al. (2003). That is, the sample $\mathbf{x}_i$ is assumed to be generated by:

$$\mathbf{x}_i = \mathbf{D}\mathbf{s}_i + \varepsilon \tag{5.16}$$

where ε is white Gaussian noise. The prior distribution of $\mathbf{s}_i$ is assumed to follow the Laplacian distribution:

$$p(\mathbf{S}|\lambda) \propto \exp\bigg(-\lambda \|\mathbf{S}\|_1\bigg) \tag{5.17}$$

where λ is a hyperparameter. This prior has been shown to encourage sparsity, which is desirable in many situations, because of its heavy tails and sharp peak. Given the prior, Bayes' theorem can now be used to express the posterior distribution for $\mathbf{S}$ as the product of the prior distribution and the likelihood function:

$$p(\mathbf{S}|\mathbf{X},\mathbf{D},\lambda) \propto p(\mathbf{S}|\lambda)p(\mathbf{X}|\mathbf{S},\mathbf{D}) \tag{5.18}$$

where $p(\mathbf{X}|\mathbf{S},\mathbf{D}) = \|\mathbf{X} - \mathbf{DS}\|_F^2$ is the likelihood function, when viewed as a function of $\mathbf{S}$. Then, maximum a posteriori (MAP) estimation of $\mathbf{S}$ can be formulated as follows:

$$\begin{aligned} \mathbf{S}^\star &= \arg\max_{\mathbf{S}} p(\mathbf{S}|\mathbf{X},\mathbf{D},\lambda) \\ &= -\arg\min_{\mathbf{S}}[-\log p(\mathbf{X}|\mathbf{S},\mathbf{D}) - \log p(\mathbf{S}|\lambda)] \\ &= \arg\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 \end{aligned} \tag{5.19}$$

Based on the discussion above, it is straightforward to show the Bayesian interpratation of the proposed model, Eq. (5.14, in this chapter. In particular, for the sake of simplicity, let's assume that $\mathbf{W}$ is fixed, and thus removed from Eq. (5.14). Also, $\mathbf{A_W}$ is denoted as $\mathbf{A}$. Then, Eq. (5.14) becomes

$$\min_{\mathbf{D},\mathbf{S}} \frac{1}{2} \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{SA}\|_1 \tag{5.20}$$

Following the same principle in Eq. (5.19), we can write:

$$\begin{aligned} \mathbf{S}^\star &= \arg\max_{\mathbf{S}} p(\mathbf{S}|\mathbf{X}, \mathbf{D}, \mathbf{A}, \lambda) \\ &= -\arg\min_{\mathbf{S}} [-\log p(\mathbf{X}|\mathbf{S}, \mathbf{D}) - \log p(\mathbf{S}|\lambda, \mathbf{A})] \\ &= \arg\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{SA}\|_1 \end{aligned} \tag{5.21}$$

The only difference of Eq. (5.21) compared to Eq. (5.19) is the additional graph matrix $\mathbf{A}$: Eq. (5.20) considers underlying data structure during the model learning, while Eq. (5.15) is not.

5.1.5 Cross-view Matching

After learning the dictionary $\mathbf{D}$ using the unlabelled training data $\mathbf{X}$, given a pair of test samples $\mathbf{x}_i^a$ and $\mathbf{x}_i^b$, first their collaborative representations $\mathbf{s}_i^{a*}$ and $\mathbf{s}_i^{b*}$ are computed by solving the following problems:

$$\mathbf{s}_i^{a*} = \arg min_{\mathbf{s}_i^a} \|\mathbf{x}_i^a - \mathbf{D}\mathbf{s}_i^a\|_F^2 + \lambda \|\mathbf{s}_i^a\|_2^2 \tag{5.22}$$

$$\mathbf{s}_i^{b*} = \arg min_{\mathbf{s}_i^b} \|\mathbf{x}_i^b - \mathbf{D}\mathbf{s}_i^b\|_F^2 + \lambda \|\mathbf{s}_i^b\|_2^2 \tag{5.23}$$

These are standard l_2-norm regularised least squares problems with closed-form solutions: $\mathbf{s}_i^{a*} = \mathbf{P}\mathbf{x}_i^a$ and $\mathbf{s}_i^{b*} = \mathbf{P}\mathbf{x}_i^b$, where $\mathbf{P} = (\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top$. Then, after obtaining $\mathbf{s}_i^{a*}$ and $\mathbf{s}_i^{b*}$ their cosine distance is used to measure the visual similarity for ReID. Hence, the proposed model is very efficient in testing.

5.1.6 Extension to Supervised ReID

Although the proposed model in this work is tailored for unsupervised ReID, it can be easily extended if labelled cross-view pairs become available. More specifically, the label information can be encoded in the graph $\mathbf{W}$. That is, instead of learning $\mathbf{W}$, it is now fixed so that if the corresponding cross-view pair (i, j) is labelled as containing the same person, $\mathbf{W}_{i,j}$ is set to 1, otherwise it is set to 0. This essentially gives thus the ideal graph and the relaxed visual similarity constraint becomes a more stringent identity constraint which requires that people of the same identity to be close in the learned attribute space and vice versa. Note that in this case, the learning graph is not performed, that is, constraints regarding to **W** are also removed in Eq. (5.14) – only learning of **D** and **S**. That is,

$$\min_{\mathbf{D},\mathbf{S}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda_1 \|\mathbf{S}\mathbf{A}_{\mathbf{W}}\|_1\| \qquad s.t. \quad \|\mathbf{d}_i\|_2^2 \leq 1 \tag{5.24}$$

In fact, the empirical experiments show that after constructing the graph with supervised information, there is no significant performance difference between Eq. (5.24) and Eq. (5.14).

5.2 Experiments and Evaluations

5.2.1 Datasets and Settings

Datasets. Four widely used benchmark datasets are used for the experiments. *VIPeR* (Gray et al., 2007) contains 632 image pairs of people captured outdoor from two non-overlapping camera views. Following the standard setting which is single-shot i.e., one image per person per view, the dataset is randomly split into two sets of 316 image pairs, one for training and the other for testing. For the test set, all images from one view is used as the gallery set and the others as probe set. Note that the identities of people in training set is completely different from that of the test set. The results for all evaluations were obtained by averaging over 10 splits. *PRID* (Hirzer et al., 2011) is different from the other available datasets in that the gallery and probe sets have different numbers of people. There are two version of it: multi-shot and single-shot (basically first frame is taken from multi-shot). In the experiments, the single-shot version of the dataset is used as in (Giuseppe et al., 2014; Hirzer et al., 2012; Paisitkriangkrai et al., 2015), while multi-shot version is used in video-based ReID McLaughlin et al. (2016).

Specifically, out of the 749 people captured in two camera views, only 200 people appear in both views. In each data split, 100 out of that 200 people are chosen randomly for training, while the remaining 100 of one view are used as the probe set, and the remaining 649 people's images of the other view are used as gallery, which thus includes the 100 people in the probe set. Experiments are carried out on the same 10 splits as in (Giuseppe et al., 2014; Hirzer et al., 2012) with the average results reported. *CUHK01* (Li et al., 2012) consists of 971 people with two images per person per camera view i.e. multi-shot. The standard setting is used as in (Li et al., 2012): 486 persons for training, while 485 persons for test. *CUHK03* (Li et al., 2014b) contains 13,164 images of 1,467 people. Two versions exist which differ in whether the images were obtained by manual cropping or automatically by applying the DPM person detector (Felzenszwalb et al., 2010). The detector-generated images are used as they reflect better the real-world application scenarios for testing the robustness of the proposed model in this work against outliers. There are in total six camera views but each person is observed in only two out of the six views, and has 4.8 images on average for each view. The same setting and random splits are used as in (Li et al., 2014b) with a single-shot setting: for the test set 100 people are randomly selected with two images each, whilst images of the remaining people are used for training. Note that out of the four datasets, CHUK03 is much bigger than the other three in terms of both the number of identities and the number of images in the training set. The summary of these datasets is given in Table 2.2, and examples are shown in Figure 2.18.

Settings. *Features*: The features introduced in (Giuseppe et al., 2014) are adopted. Each image is scaled to 128×48 in all datasets, and then histogram-based image descriptors (window-based representation) are computed consisting of three types: (1) Colour histogram using HS, RGB, and Lab colour spaces (2880-D colour vector), (2) HOG (1040-D), and (3) LBP (1218-D) (Ahonen et al., 2004). The final image feature vector, 5138-D, is obtained as the concatenation of these three types of features. *Evaluation metrics*: the Cumulative Matching Characteristics (CMC) curves are obtained as an evaluation metric. Matching accuracies at Rank 1 in all datasets and the full CMC curves for VIPeR and CUHK01 are reported. *Parameter settings*: There are a number of parameters in the proposed model in this work. As an unsupervised learning method, there are no other means but setting them manually. For the dictionary size k, it is not tuned carefully and it is set to 256 for the two small datasets VIPeR and PRID, and 512 for the larger CUHK01 and CUHK03 dataset. Its effects on the performance will be discussed later. In the objective function (Eq. (5.14)), there are two weights λ_1 and λ_2 for the two regularisation terms respectively. As explained in Section 5.1.3, λ_2 is set automatically using Eq. (5.13) in the ADMM algorithm, whilst for λ_1 it is set to 1 throughout, as it is found that the algorithm is insensitive to its value. Similarly for the initial construction of graph $\mathbf{G}$, a KNN (K-nearest neighbour) graph is used with cosine distance and $K = 5$ for all datasets.

5.2.2 Evaluation of Unsupervised Learning based ReID

Compared methods. Under this setting, the proposed approach (L_1Graph) is compared with state-of-the-art unsupervised alternatives which fall into four categories:

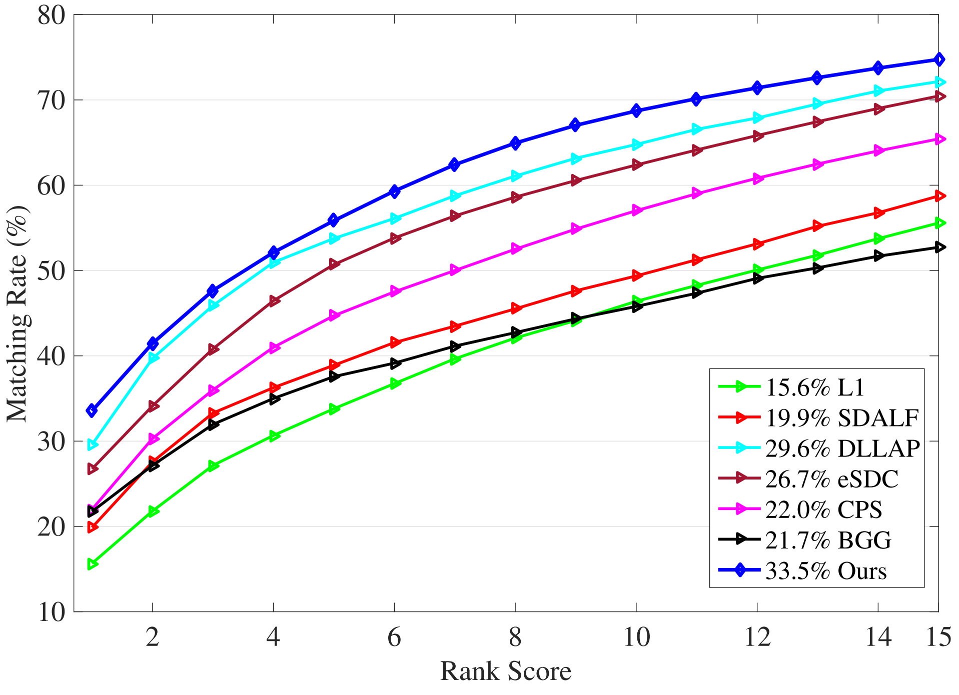
1) *The hand-crafted feature-based methods*: SDALF (Farenzena et al., 2010) and CPS (Cheng et al., 2011). Both methods use human structure while designing the feature representation. For example, CPS detects body parts, extract features from each part, and concatenate the features to obtain high dimensional feature representation. After that any distance measurement can be used over the representations.

2) *The saliency learning-based methods*: eSDC (Zhao et al., 2013a) and GTS (Wang et al., 2014). Bot methods try to make use of saliencyt information to handle misalignment problem in pedestrian images.

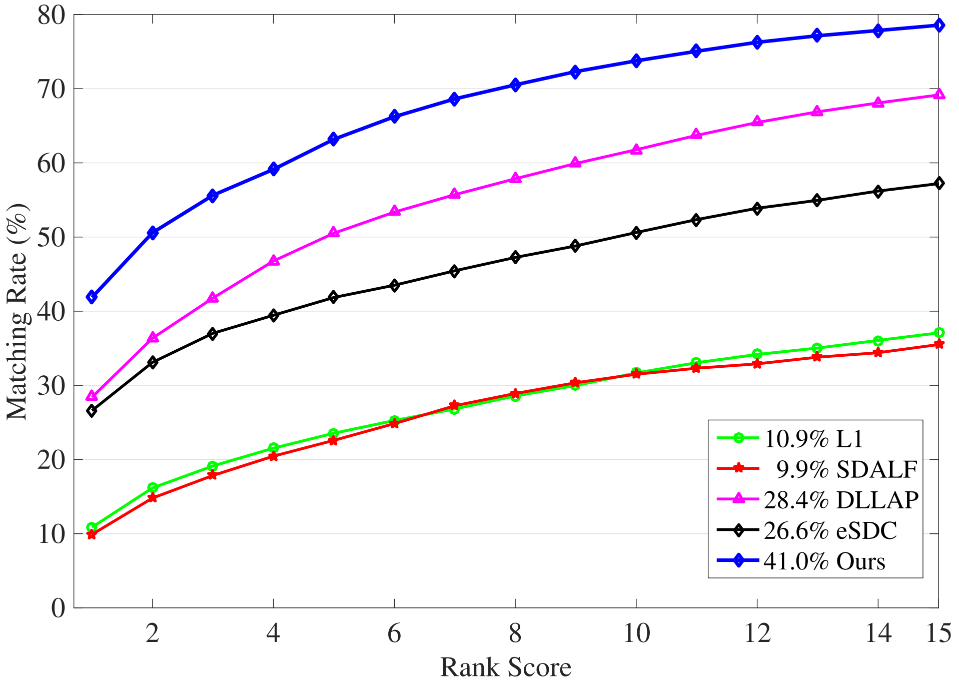
4) *The codebook learning-based method*: BGG (Zheng et al., 2015a) uses bag-of-visual words to learn a feature representation (Csurka et al., 2004).

3) *The dictionary learning-based methods*: DLLAP (Kodirov et al., 2015), and UCDTL (Peng et al., 2016) which use the same 5138-D features for fair comparison. Both methods including the proposed approach in this thesis attempt to learn feature representation out of low-level features. The learned features are then assumed to be discriminative across non-overlapping camera views. Note that DLLAP and UCDTL use a *conventional* graph regularisation, thus prone to noises and outliers, whereas this thesis uses robust graph regularisation with combination to dictionary learning as well as graph learning.

Results. Table 5.1 compares the results of the L_1Graph against the six alternatives and a non-learning l_1 distance based baseline. From Table 5.1, the following observations can be made: (1) The proposed robust graph regularised dictionary learning model (L_1Graph) outperforms all existing unsupervised methods on all four datasets, and often by a large margin. (2) The margin is in general bigger on the two larger datasets CUHK01 and CUHK03, which indicates that L_1Graph can benefit more from larger unlabelled training data. (3) Among the alternatives, the dictionary learning based methods such as DLLAP and UCDTL are the most competitive. Thanks to the introduced two novel components: robust graph regularisation and joint graph and dictionary



(a) VIPeR dataset



(b) CUHK01 dataset

Figure 5.2: CMC curves for VIPeR and CUHK01. The curves correspond to Table 5.1. Legends states the rank 1 performance.

Datasets	*VIPeR*	*PRID*	*CUHK01*	*CUHK03*
l_1	15.6	13.9	10.9	12.5
SDALF (Farenzena et al., 2010)	19.9	16.3	9.9	4.9
eSDC (Zhao et al., 2013a)	26.7	-	26.6	7.7
CPS (Cheng et al., 2011)	22.0	-	-	-
GTS (Wang et al., 2014)	25.2	-	-	-
BGG (Zheng et al., 2015a)	21.7	-	-	18.9
DLLAP (Kodirov et al., 2015)	29.6	21.1	28.4	22.3
UCDTL (Peng et al., 2016)	31.5	24.2	27.1	-
L_1Graph	**33.5**	**25.0**	**41.0**	**30.4**

Table 5.1: Unsupervised ReID results measured in Rank-1 matching accuracy (%) on VIPeR, PRID, CUHK01, CUHK03, where '-' denotes no reported result.

learning. This result also suggests that learning a low-dimensional latent attribute representation is more suited for unsupervised ReID than the alternative models. In particular, the difference between L_1Graph and l_1 is large which means that matching people is made much easier in this learned discriminative subspace with less than one tenth of the original dimensions. Furthermore, Figure 5.2, which corresponds to Table 5.1 shows the CMC curves of L_1Graph for VIPeR and CUHK01 with comparison to the state-of-the-arts. Compared methods for VIPeR: l_1 distance, SDALF, DLLAP, eSDC, CPS, BGG, and for CUHK01: l_1 distance, SDALF, DLLAP, eSDC. Overall, as it can be seen from the Figure 5.2, L_1Graph outperforms all its competitors in all ranks. The advantage of L_1Graph's computational efficiency over other methods will be discussed later.

5.2.3 Evaluation of Supervised Learning based ReID

Compared methods. Since the performance of different existing methods on different datasets often vary drastically[3], the best methods were chosen for each dataset separately to better reflect the state-of-the-art. All methods are published in the last three years. Note that multi-feature fusion-based methods are separated from single feature or deep models as typically any method can benefit from multi-feature fusion. As mentioned in Section 5.1.6, L_1Graph can also operate

[3] For example, deep learning based methods often perform stronger on the large datasets than the small ones due to the need for large training data.

in the supervised mode; denoted as L_1Graph_sup, this can be considered as the upper bound of L_1Graph's performance under the unsupervised setting when the graph is learned perfectly.

Results. The following key findings are observed from Table 5.2: (1) The gap between L_1Graph_un and L_1Graph_sup is moderate. This indicates that L_1Graph is very effective and the performance of the unsupervised model is not far off from its upper bound. (2) On the two smaller datasets, VIPeR and PRID, L_1Graph is very competitive under the supervised setting: on VIPeR it beats all single feature-based methods, and on PRID it outperforms all existing supervised methods, often significantly. Even L_1Graph outperforms some very recent supervised models. (3) On the two larger datasets CUHK01 and CUHK03 (with detected person images), the gap between L_1Graph and the state-of-the-art begins to appear. L_1Graph (both in supervised and unsupervised cases) remains competitive on CUHK01, but on CUHK03, the gap is big, in particular to L_1Graph under unsupervised setting. This is expected: with over 10,000 labelled training images from 1,367 people, an unsupervised model cannot compete with a supervised one, especially those based on deep learning. However, it is noteworthy to point out that in practice collecting hundreds of labelled training samples is very difficult and collecting thousands would be near impossible across even just a handful of camera views. Moreover, current large-scale datasets, were collected from very constrained environments under about 2-10 cameras Zheng et al. (2015a).

5.2.4 Further Analysis

Ablation study. The proposed L_1Graph has two key components and to see the impact of each the full model it is compared with various striped-down versions of the model under the unsupervised setting: (1) L_1Graph_*DL* – without graph regularisation which is the same as conventional dictionary learning; (2) L_1Graph_l_2 – the graph is fixed and l_2-norm is used for graph regularisation; (3) L_1Graph_l_2_graph – the graph is learned and l_2-norm is used for graph regularisation; (4) L_1Graph_l_1 – the graph is fixed and l_1-norm is used for graph regularisation; (5) L_1Graph_full – the full proposed model in which the graph is learned and l_1-norm is used for graph regularisation. Table 5.3 shows that both using robust l_1-norm graph regularisation and joint graph and dictionary learning contribute positively toward the final performance. The result (comparing L_1Graph_*DL* with the other models) also shows that adding a graph regularisation term to learn cross-view discriminative information in general is critical for dictionary-learning-based ReID.

Datasets	*VIPeR*		*PRID*		*CUHK01*		*CUHK03*	
	Reference	Rank 1	Reference	Rank 1	Reference	Rank 1	Reference	Rank 1
Single-feature Methods	(Liao et al., 2015)	40.0	(Giuseppe et al., 2014)	14.5	(Zhao et al., 2014a)	34.3	(Liao et al., 2015)	46.4
	(Ahmed et al., 2015)*	34.8	(Xiong et al., 2014a)	19.7	(Ahmed et al., 2015)*	47.5	(Wei et al., 2017)	**80.6***
	(Wei et al., 2017)*	**50.2**	(Peng et al., 2016)	24.2	(Wei et al., 2017)*	**91.2**	(Xiao et al., 2016)*	75.3
	(Liao and Li, 2015)	40.7	(Kodirov et al., 2015)	25.2	(Li et al., 2014b)*	29.4	(Liao and Li, 2015)	51.2
	(Chen et al., 2015)	36.1	(Roth et al., 2014)	16.0	(Li et al., 2014b)*	27.8	(Li et al., 2014b)	19.9
	(Shi et al., 2015)	40.9	(Su et al., 2015)	18.0	(Liao et al., 2015)	63.5	(Shi et al., 2015)	52.1
	(Zheng et al., 2015b)	30.2	(Liao and Li, 2015)	12.3	(Liao and Li, 2015)	64.2	(Ustinova et al., 2015)*	59.2
Multi-feature Fusion	(Paisitkriangkrai et al., 2015)	45.9	(Paisitkriangkrai et al., 2015)	17.9	(Paisitkriangkrai et al., 2015)	53.4	–	–
L_1**Graph_un**	33.5		25.0		41.0		30.4	
L_1**Graph_sup**	41.5		**30.1**		50.1		39.0	

Table 5.2: Comparison with state-of-the-art supervised learning-based methods. '*' indicates deep methods.

Methods	L_1Graph_*DL*	L_1Graph_l_2	L_1Graph_l_2_graph	L_1Graph_l_1	L_1Graph_full
VIPeR	19.6	29.4	30.1	32.0	33.5
CUHK01	17.4	36.9	37.5	38.7	41.0

Table 5.3: The contributions of individual model components.

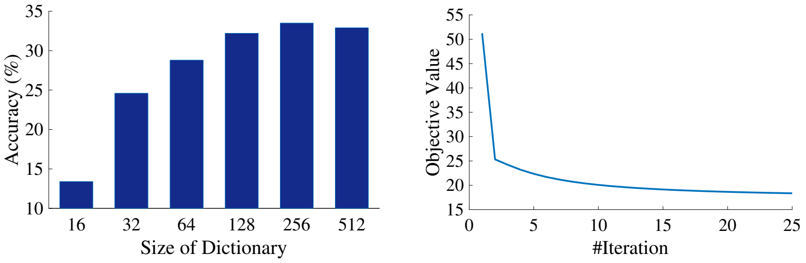


Figure 5.3: (Left) Rank 1 accuracies with different dictionary sizes on VIPeR dataset; (Right) Objective function value with respect to the number of iterations on CUHK01.

Effect of dictionary size and convergence analysis. The only parameter that needs to be tuned for each dataset is the dictionary size. Figure 5.3(Left) shows that when the size is over 100, its effect is small. Furthermore, Figure 5.3(Right) shows the proposed method converges rapidly. Although there is no theoretical proof, convergence is observed in all the experiments within 25 iterations.

Running cost. All the experiments were conducted in MATLAB on a PC with two 3.40 GHz CPUs and 16G RAM. The training of the model on VIPeR takes 178.3 seconds but during test it is very efficient: once the 5138-D features are extracted, it takes only 0.01 second to match one probe image against 316 images from the gallery. Table 5.4 compares the running time of feature extraction and matching during test time against a number of alternative unsupervised methods (whose source codes in MATLAB are publicly available). It is clear that L_1Graph is often a few magnitudes faster than its competitors.

Stage	SDALF	eSDC	BGG	L_1Graph
Feature Extraction (s)	2.92	0.76	0.62	0.03
Matching (s)	550.80	9.7	0.44	0.01

Table 5.4: Average testing time of different methods on VIPeR

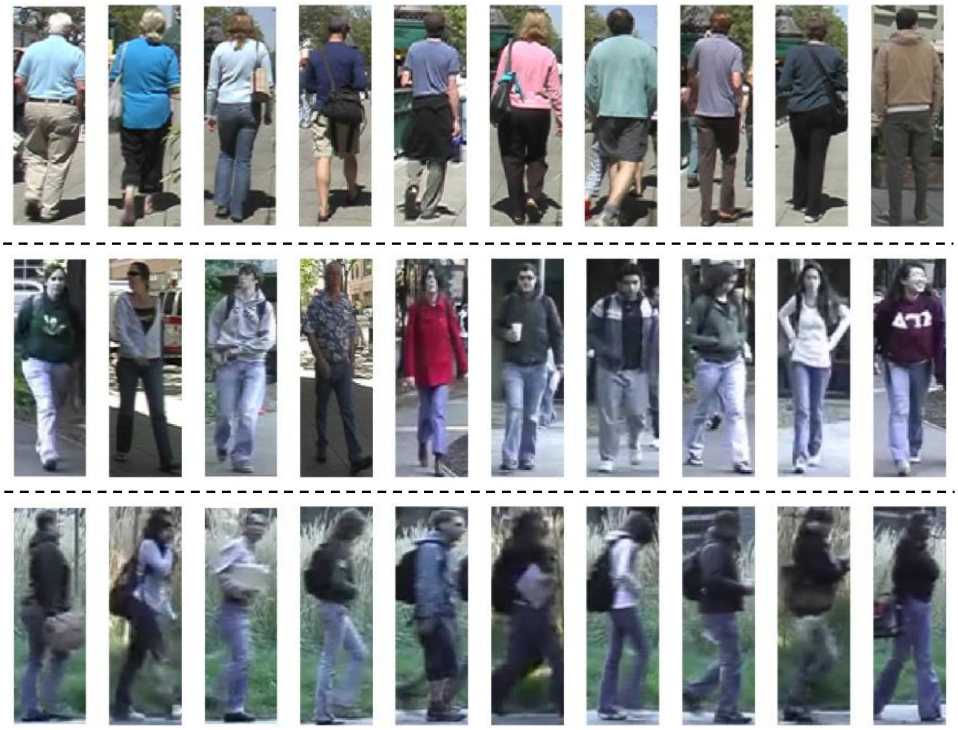
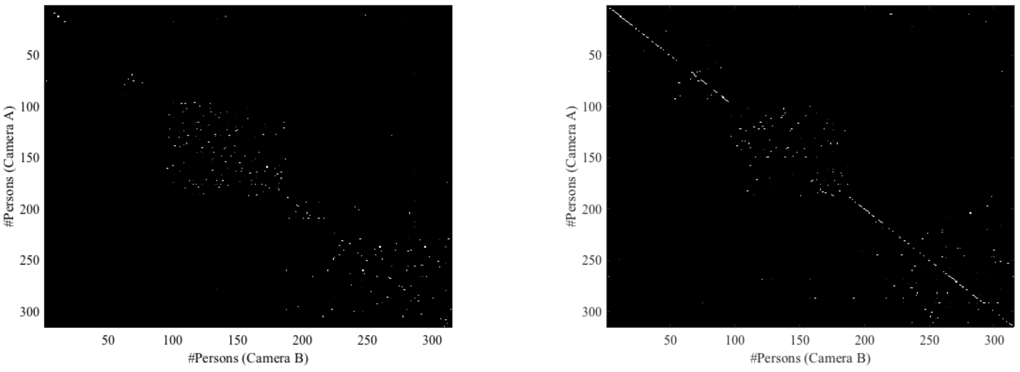


Figure 5.4: Dictionary visualisation. Example images whose features are closest to a certain basis. Each row corresponds to a certain dictionary atom.

Dictionary visualisation. Visualisation of the learned dictionary is given in Figure 5.4. The dictionary basis is visualised via exemplar images from dataset. Each row in Figure shows five example images whose features are closest to a certain basis. The exemplars suggest that the learned dictionary discovers certain common structures across multiple persons. For example, the basis in the first row depicts 'back' view (latent attribute), and the middle row images has 'wearing jeans' (latent attribute), and 'front' view (latent attribute) of persons. Specifically, d-dimensional feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ are extracted for all given images. Then, a basis is chosen which is also a d-dimensional vector $\mathbf{d}_i \in \mathbb{R}^d$. After that, the distance between the feature vectors and the basis is computed, and top 5 nearest feature vectors are chosen in terms of distance. The images are retrieved according to the indices of the top 5 feature vectors are shown in Figure 5.4).

5.3 Conclusion

This chapter presented a novel unsupervised ReID model based on dictionary learning. The key contributions: (1) the introduction of a robust l_1-norm graph regularisation term which is robust



(a) Before

(b) After

Figure 5.5: An illustration of the graph **W** before and after learning process. Ideally, **W** should be identiy matrix. VIPeR dataset is used: 316 persons from Camera A and 316 persons from Camera B.

against outliers and noise abundant in person ReID. This robust regularisation is then integrated into the standard dictionary formulation to learn latent attributes (as dictionary atoms) which are invariant across different camera views. In other words, latent attributes capture cross-view discriminative information which are essential for finding a correct match of the person across non-overlapping cameras. Although the proposed robust graph regularisation is advantageous over the conventional robust regularisation according to the extensive experiments conducted in this chapter, it still relies on the graph constructed from the original data, meaning the graph itself may contain some connected edges which do not reveal actual data graph topology. To this end, (2) a joint graph and dictionary learning algorithm is developed which further improves the ability of the proposed model to deal with outlying samples. In addition, Bayesian interpretation of the proposed regularisation is discussed. Extensive experiments on four benchmark datasets show that the proposed method significantly outperforms existing unsupervised methods in terms of cumulative matching curve and computational cost. Also, the proposed method is compared to supervised models to show how far the unsupervised models are, showing the supervised models are far superior than the unsupervised ones. Nevertheless, it needs to be stressed that considering scalability unsupervised methods are more important given hundreds of cameras, while obtaining annotated information for supervised models for hundreds of cameras are prohibitively expensive if impossible.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has presented three variants of cross-class transfer learning for visual recognition. In particular, transductive, inductive and unsupervised settings are investigated and explored. Specifically,

1. In Chapter 3, Transductive cross-class transfer learning framework is formulated based on regularised dictionary learning for taking into account available target data without labels. Compared with most existing CCTL methods that perform naive transfer, the proposed model is essentially an unsupervised domain adaptation model which learns an embedding function from a visual space to a semantic space using both labelled source and unlabelled target data. Extensive comparative evaluations validate the advantages of the model over the state-of-the-arts.
2. In Chapter 4, Inductive cross-class transfer learning method, SAE, is presented, which is more practical. That is, the proposed model is characterised with the ability of learning more generalisable knowledge without accessing to unlabelled target data. The SAE model uses very simple and computationally fast linear embedding function and introduce an additional reconstruction objective function for learning a more generalisable embedding function. Thesis demonstrates through extensive experiments that this new SAE model outperforms existing CCTL models on six benchmarks. Furthermore, the model is further extended to a deeper SAE by introducing more layers.

3. In Chapter 5, Unsupervised cross-class transfer learning method is proposed for instance recognition. This model is applied to person re-identification which is an instance of the verification problem. Different from most contemporary person re-identification methods, the proposed method does not require labelled data, yet learn an identity discriminative patterns for recognising people across non-overlapping cameras at different locations and time. The key contributions are the introduction of a robust l_1-norm graph regularisation term in the dictionary learning formulation so that cross-view discriminative information can be learned. In addition, a joint graph and dictionary learning algorithm is developed which further improves the ability of the proposed model to deal with outlying samples abundant in person ReID data. Extensive experiments on four benchmark datasets show that the proposed method significantly outperforms existing unsupervised methods.

Although the newly proposed methods have explored several issues and challenges in visual recognition, they need to be further explored and investigated from different directions and perspectives, and a few of them are discussed below.

6.2 Future Work

The future research directions beyond the proposed methods in this thesis are summarised as follows:

1. (Chapter 3) **Transductive cross-class transfer learning**: The proposed model in this setting is based on dictionary learning. There are several ways of extending current approach. Firstly, this chapter proposed a method that learns a semantic dictionary that cares only semantic attributes. In practise, however, since the images is often occluded or have background clutter, true semantic dictionary could not be obtained. To alleviate this, latent attributes can also be discovered along with the semantic dictionary simultaneously (Yang et al., 2014b). In this case, latent attributes try to capture occlusion or different patterns, resulting in better semantic dictionary. Secondly, current approach is linear and cannot capture nonlinearity of the data. Hence, non-linear version could also be considered for further extension (Van Nguyen et al., 2013).
2. (Chapter 4) **Inductive cross-class transfer learning**: By only relying on labelled data, yet achieving a good generalisation model is ultimate goal in this setting. However, since visual data is very high dimensional, capturing various characteristics of real world visual

data by only relying on training set is very challenging. To this end, recently deep generative models gain significant attention. So, one possible extension could be integrating the current model into generative models. Also, it is interesting to see the other applications of the proposed approach such as supervised clustering (Law et al., 2016), and supervised learning-based person re-identification methods (Ahmed et al., 2015). Furthermore, the proposed regularisations in Chapter 3 can be explored in combination with the SAE to see whether they are effective in the SAE. The proposed deep SAE also needs further investigation. Extending deep SAE to learn en embedding function an end-to-end manner meaning features and semantic embedding are learned jointly.

3. (Chapter 5) **Unsupervised cross-class transfer learning**: Discovering discriminative patterns in an unsupervised manner remains an open issue although large strides have been made recently. Indeed, more further research effort is still needed. Two lines of extension work can be considered with regards to the proposed approach. Firstly, current model can be extended to semi-supervised learning paradigm (Kingma et al., 2014) in which some partial labelled data is assumed to be available. The labels could be not only IDs of persons, but also semantic information such as attributes or pose information. Also, since deep generative models have significant success (Goodfellow et al., 2016), it is desirable to see the combination of the proposed approach and deep learning generative models (Raina et al., 2004).

Appendices

Appendix A

Bartels-Stewart Algorithm

Solution of equation Eq. (A.1) was presented in (Bartels and Stewart, 1972).

$$\mathbf{AX} + \mathbf{XB} = \mathbf{C} \tag{A.1}$$

where **A** of size $m \times m$, **B** of size $n \times n$, and **C** of size $m \times m$ are matrices with real elements. The solution is based on Schur decomposition. Bartesl-Stewart algortihm consists of mainly four steps, and they are as follows:

1) Transforming **A** and **B** to real Schur form:

$$\mathbf{S_A} = \mathbf{Q}^\top \mathbf{AQ} \tag{A.2}$$

$$\mathbf{S_B} = \mathbf{Q}^\top \mathbf{BQ} \tag{A.3}$$

   where **Q** and **P** are orthogonal matrices, and $\mathbf{S}_A$ and $\mathbf{S}_B$ are in real Schur form.

2) Updating **C** with respect to the two Schur decompositions:

$$\tilde{\mathbf{C}} = \mathbf{Q}^\top \mathbf{CP} \tag{A.4}$$

3) Solving the resulting reduced triangular matrix:

$$\mathbf{S_A}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}\mathbf{S_B} = \tilde{\mathbf{C}} \tag{A.5}$$

4) Transforming the obtained solution back to the original coordinate system.

$$\mathbf{X} = \mathbf{Q}\tilde{\mathbf{X}}\mathbf{P}^\top \tag{A.6}$$

For more infomation please refer to the original paper (Bartels and Stewart, 1972).

Appendix B

Label Propagation

A main idea of label propagation (LP for short) is that data points that are close to have similar labels. *Definition*: Let $\{(\mathbf{x}_1, \mathbf{z}_l), \ldots, (\mathbf{x}_l, \mathbf{z}_l)\}$ be labelled data, where $\mathbf{X}_L = \mathbf{x}_1, \ldots, \mathbf{x}_l$ are examples, and $\mathbf{Z}_L = \mathbf{z}_1, \ldots, \mathbf{z}_l$ are the corresponding labels. It is assumed that the number of classes C is known, and all classes are present in the labelled data. Similarly, let $\{(\mathbf{x}_{l+1}, \mathbf{z}_{l+u}), \ldots, (\mathbf{x}_{l+1}, \mathbf{z}_{l+u})\}$ be unlabelled set (unobserved), where $\mathbf{X}_U = \{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}\}$ and $\mathbf{Z}_U = \{\mathbf{z}_{l+1}, \ldots, \mathbf{z}_{l+u}\}$. Let $\mathbf{X} = \{\mathbf{X}_L, \mathbf{Z}_U\}$. The goal is to estimate $\mathbf{Z}_U$ from $\mathbf{X}$ and $\mathbf{Z}_L$ – a transductive learning setting.

Firstly, a fully connected graph where the nodes are all data examples are created using $\mathbf{X}$. The edge between any noes $\{i, j\}$ is weighted so that the closer the nodes are in local Euclidean distance, the larger the weight g_{ij}. The weights are controlled by a parameter σ:

$$g_{ij} = \exp\left(-\frac{\mathbf{d}_{ij}^2}{\sigma^2}\right) = \exp\left(\frac{\sum_{d=1}^{D}(\mathbf{x}_i^d - \mathbf{x}_j^d)^2}{\sigma^2}\right) \tag{B.1}$$

All nodes have soft labels due to the exp function that can be treated as distributions over labels. It is let that the labels of a node to propogate to all nodes through the edges. Larger edge weights allow labels to travel through easier. Define of size $(l+u) \times (l+u)$ probabilistic transition matrix $\mathbf{P}$ as follows:

$$\mathbf{P}_{ij} = \mathbf{P}(j \to i) = \frac{g_{ij}}{\sum_{k=1}^{l+u} g_{kj}} \tag{B.2}$$

where $\mathbf{P}_{ij}$ is the probability to jump from node j to i. Also, let's define $(l+u) \times C$ label matrix $\mathbf{Y}$, whose ith row representing the label probability distribution of node $\mathbf{x}_i$. The initialisation of rows of **Y** corresponding to unlabelled data examples is unimportant.

The LP algorithm (Xiaojin and Zoubin, 2002) is composed of three steps: (1) all nodes propagate their labels for one step; (2) Row-normalising **Y** to maintain the label probability interpretation; (3) This step is critical in which persistent label sources from labelled data is desired. Therefore, instead of letting the initially labelled nodes fade away, they are replenished by clamping their label distribution to $\mathbf{Y}_{ic} = \delta(\mathbf{y}_i, c)$, that is, the probability mass is concentrated on the given class. With this constant 'push' from labelled nodes, the class boundaries will be pushed through high density data filaments and settle in low density gaps.
A brief summary of the LP algorithm is as follows:

1) Propagate $\mathbf{Y} \leftarrow \mathbf{PY}$.

2) Normalise **Y** in terms of rows.

3) Clamp the labelled samples. Repeat from step 1 until **Y** converges.

For detailed explanation and convergence analysis of the LP algorithm, please see (Xiaojin and Zoubin, 2002; Fu et al., 2014).

Appendix C

Indirect Attribute Prediction

Indirect Attribute Prediction (IAP) is a classic probabilistic approach for cross-class recognition (Lampert et al., 2009). It uses attributes to transfer knowledge between classes. IAP is illustrated in Figure C.1 in which the attributes ($\{a_1, \ldots, a_M\}$) form a connecting layer between two layers of labels, one for training classes ($\{y_1, \ldots, y_K\}$), and the other for test classes ($\{z_1, \ldots, z_L\}$). The training phase of IAP is standard multi-class classificaiton. During test time, the predictions for all the training classes induce a labeling of the attribute layer, from which a labelling over the test classes can be inferred.

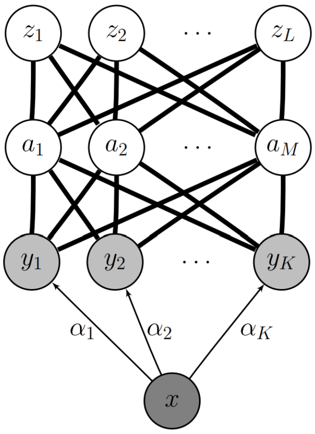


Figure C.1: Indirect attribute prediction (IAP). The figure is taken from (Lampert et al., 2009)

More formally, in the first stage, a probabilistic multi-class classifier estimating $p(y_k|x)$ is learned for all training classes y_1, ..., y_k. Then, it is assumed that there is a deterministic dependence between attributes and classes, thus $p(a_m|y)$ is et to $[a_m = a_m^y]$. Both stages yields

$$p(a_m|x) = \sum_{k=1}^{K} p(a_m|y_k)p(y_k|x) \tag{C.1}$$

From Eq. C.1, inferring the attribute posterior probabilities $p(a_m|x)$ requires only a matrix-vector multiplication (these probabilities have been used in Chapter 3). Afterwards, classifying test examples can be performed by the following equation:

$$f(x) = \underset{l=1,\ ...,\ L}{\arg\max}\, \Pi_{m=1}^{M} \frac{p(a_M^{z_l}|x)}{p(a_m^{z_l})} \tag{C.2}$$

where $p(a_m) = \frac{1}{K}\sum_{k=1}^{K} a_m^{y_k}$ is empirical means over the training classes as attribute priors.

Bibliography

M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. *European Conference on Computer Vision*, 2004.

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

D. P. Ausubel. Educational psychology: A cognitive view. *New York: Holt, Rinehart and Winston*, 1968.

V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010.

R. Bartels and G. Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 1972.

A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 2014.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

C. M. Bishop et al. Bayesian regression and classification. In *Advances in Learning Theory: Methods, Models and Applications*. Citeseer, 2003.

Y.-l. Boureau, Y. L. Cun, et al. Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems*, 2008.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

A. L. Brown and M. J. Kane. Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20(4):493–523, 1988.

M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, 2016.

K. R. Canini, M. M. Shashkov, and T. L. Griffiths. Modeling transfer learning in human categorization with the hierarchical dirichlet process. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 151–158. Omnipress, 2010.

G. Carneiro and D. Lowe. Sparse flexible models of local features. *European Conference on Computer Vision*, 2006.

J. D. N. . A. J. Caas. Psychological foundations of human learning. In *http://cmap.ihmc.us/docs/psychologicalfoundations.php*, 2010.

S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, 2016a.

W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, 2016b.

O. Chapelle, B. Schölkopf, A. Zien, et al. *Semi-supervised learning*, volume 20. MIT press Cambridge, 2006.

D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, 2011.

F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *IEEE International Conference on Computer Vision*, 2011.

C. Cortes and V. Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

G. Csurka, C. Dance, L. Fan, and J. Willamowski. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*, 2004.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine learning*, 2007.

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*, 2009.

J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*. Springer, 2014.

S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

C. Domeniconi, D. Gunopulos, and J. Peng. Large margin nearest neighbor classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, 16(4):899–909, 2005.

M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *IEEE International Conference on Computer Vision*, 2013.

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*, 2010.

A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference on Computer Vision*, 2013.

D. R. Fokkema, G. L. Sleijpen, and H. A. Van der Vorst. Jacobi–davidson style qr and qz algorithms for the reduction of matrix pencils. *SIAM journal on scientific computing*, 20(1): 94–125, 1998.

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013.

Y. Fu and L. Sigal. Semi-supervised vocabulary-informed learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*, 2014.

Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2332–2345, 2015a.

Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015b.

Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015c.

P. Genevieve, X. Chen, S. Hang, and H. James. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

L. Giuseppe, M. Iacopo, and D. B. Alberto. Matching people across camera views using kernel canonical correlation analysis. In *ACM International Conference on Distributed Smart Cameras*, 2014.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

S. Gong, M. Cristani, S. Yan, and C. Loy. *Person Re-Identification*. Springer, 2014.

I. Goodfellow. Deep learning of representations and its application to computer vision. *Thesis*, 2015.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

K. Grauman and B. Leibe. *Visual Object Recognition*. Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool Publishers, 2011. URL `https://books.google.co.uk/books?id=lAQGBvdm3UsC`.

D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

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*, 2007.

M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, 2009.

H. Guo, Z. Jiang, and L. S. Davis. Discriminative dictionary learning with pairwise constraints. In *Asian Conference on Computer Vision*, 2012.

Y. Guo, G. Ding, X. Jin, and J. Wang. Transductive zero-shot recognition via shared model space learning. In *AAAI Conference on Artificial Intelligence*, 2016a.

Y. Guo, G. Ding, Y. Wang, and X. Jin. Active learning with cross-class knowledge transfer. In *aaai*, 2016b.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

F. Hehe, Z. Liang, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. In *https://arxiv.org/abs/1705.10444*, 2017.

A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011.

M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*. Springer, 2012.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970.

E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 2015.

K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in neural information processing systems*, pages 609–616, 2007.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.

D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, 2014.

B. Jiang, C. Ding, and J. Tang. Graph-laplacian pca: Closed-form solution and robustness. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014.

K. Kenneth, M.Joseph, R. Bhaskar, E. Kjersti, L. Te-Won, and S. Terrence. Dictionary learning algorithms for sparse representation. *Neural Computing*, 15(2):349–396, 2003.

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. \ell_1 trend filtering. *SIAM review*, 51(2): 339–360, 2009.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 2014.

G. Koch. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015.

E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *British Machine Vision Conference*, 2015.

S. Kotsiantis. Supervised machine learning: a review of classification techniques. *Informatica*, 31:249–268, 2007.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *cvpr*, 2009.

M. T. Law, Y. Yu, M. Cord, and E. P. Xing. Closed-form training of mahalanobis distance for supervised clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3909–3917, 2016.

R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *British Machine Vision Conference*, 2012.

A. Lazaridou, E. Bruni, and M. Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Annual Meeting of the Association for Computational Linguistics*, 2014.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, 2006.

J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *IEEE International Conference on Computer Vision*, 2015.

A. Li, L. Liu, and S. Yan. *Person Re-identification by Attribute-Assisted Clothes Appearance*. Person Re-Identification, 2014a.

W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, 2012.

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*, 2014b.

S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *IEEE International Conference on Computer Vision*, 2015.

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*, 2015.

Y. Liao, Y. Wang, and Y. Liu. Graph regularized auto-encoders for image representation. *IEEE Transactions on Image Processing*, 26(6):2839–2852, 2017.

C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *European Conference on Computer Vision Workshop*, 2012.

J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011a.

J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011b.

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*, 2014.

D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, 1999.

Y. Lu. Unsupervised learning on neural network outputs: with application in zero-shot learning. *arXiv preprint arXiv:1506.00990*, 2015.

X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 2017.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 2010.

A. Margolis. A literature review of domain adaptation with unlabeled data. *Technical report*, 2011.

N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

T. M. Mitchell et al. *Machine learning. WCB*. McGraw-Hill Boston, MA:, 1997.

M. Nejati, S. Samavi, N. Karimi, S. M. R. Soroushmehr, and K. Najarian. Boosted dictionary learning for image compression. *IEEE Transactions on Image Processing*, 25(10):4900–4915, 2016.

J. Ni, Q. Qiu, and R. Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

F. Nie, H. Wang, H. Huang, and C. Ding. Unsupervised and semi-supervised learning via l_1-norm graph. In *IEEE International Conference on Computer Vision*, 2011.

F. Nie, X. Wang, and H. Huang. Clustering and projected clustering with adaptive neighbors. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

M. Ozeki and T. Okatani. Understanding convolutional neural networks in terms of category-level attributes. In *Asian Conference on Computer Vision*, 2014.

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*, 2015.

S. J. Pan and Q. Yang. A survey on transfer learning. *Data Mining and Knowledge Discovery*, 22(10):1345–1359, 2010.

D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, 2011.

V. M. Patel, Y.-C. Chen, R. Chellappa, and P. J. Phillips. Dictionaries for image and video-based face recognition. *Optical Society of America*, 31(5):1090–1103, 2014.

V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

M. Perrot and A. Habrard. Regressive virtual metric learning. In *Advances in Neural Information Processing Systems*, 2015.

B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010.

R. Raina, Y. Shen, A. Mccallum, and A. Y. Ng. Classification with hybrid generative/discriminative models. In *Advances in Neural Information Processing Systems*, 2004.

N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon. Collaborative filtering with graph information: Consistency and scalable methods. In *Advances in Neural Information Processing Systems*, 2015.

S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016a.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International Conference on Machine learning*, 2016b.

M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where and why? semantic relatedness for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine learning*, 2015.

P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. *Mahalanobis distance learning for person re-identification*. Person Re-Identification, 2014.

S. Ruder. Transfer learning - machine learning's next frontier. *http://knowledgeofficer.com/knowledge*, 2017.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

A. Schumann, S. Gong, and T. Schuchert. Deep learning prototype domains for person re-identification. In *IEEE International Conference on Image Processing*, 2017.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *http://arxiv.org/abs/1312.6229*, 2013.

H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, and S. Z. Li. Constrained deep metric learning for person re-identification. *arXiv preprint arXiv:1511.07545*, 2015.

Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, 2013.

K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *IEEE International Conference on Computer Vision*, 2015.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2017.

M. Szummer and R. W. Picard. Temporal texture modeling. In *IEEE International Conference on Image Processing*, 1996.

M. Thomas, V. Jakob, P. Florent, and C. Gabriela. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision*, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996.

E. Ustinova, Y. Ganin, and V. S. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *CoRR*, 2015. URL `http://arxiv.org/abs/1512.05300`.

H. Van Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa. Design of non-linear kernel dictionaries for object recognition. *IEEE Transactions on Image Processing*, 22(12):5123–5135, 2013.

R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2):29, 2013.

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*, 2011.

H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, 2013.

H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *British Machine Vision Conference*, 2014.

H. Wang, X. Zhu, T. Xiang, and S. Gong. Towards unsupervised open-set person re-identification. In *IEEE International Conference on Image Processing*, 2016a.

K. Wang, L. Lin, W. Zuo, S. Gu, and L. Zhang. Dictionary pair classifier driven convolutional neural networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016b.

X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *IEEE International Conference on Computer Vision*, 2013.

L. Wei, Z. Xiatian, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, 2017.

K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, 2005.

J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

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*, 2016.

Z. Xiaojin and G. Zoubin. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University*, 2002.

F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, 2014a.

F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *eccv*, 2014b.

X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, 2016.

F. Yang, Z. Jiang, and L. S. Davis. Online discriminative dictionary learning for visual tracking. In *IEEE Winter Conference on Applications of Computer Vision*, 2014a.

J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, 2012.

M. Yang, D. Dai, L. Shen, and L. Van Gool. Latent dictionary learning for sparse representation based classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014b.

Y. Yang and T. Hospedales. A unified perspective on multi-domain and multi-task learning. In *International Conference on Learning Representations*, 2015.

D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *arXiv e-prints*, 2014.

F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S. Chang. Designing category-level attributes for discriminative visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013a.

R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision*, 2013b.

R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013c.

R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014a.

R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identfiation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014b.

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*, 2015a.

L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015b.

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

M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

W. Zheng, S. Gong, and T. Xiang. Person re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

X. Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4, 2005.

X. Zhu. Semantic structure discovery in surveillance videos. *Thesis*, 2015.