THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Discovering Visual Attributes from Image and Video Data

Liangchen Liu

B. Eng., M. Eng.

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in 2016

School of Information Technology and Electrical Engineering

Abstract

Visual attributes have recently attracted great attention from the computer vision community. Their advantages include being: (1) shared across classes, (2) human understandable, and (3) machine computable. Despite these positives, the attribute features in current works are not discriminative enough to be considered as strong features for traditional classification tasks or novel applications. This problem hinders their broad usage. There are several works that primarily focus on increasing the discriminative power of visual attributes. Unfortunately, in most cases, the feature descriptors become hard to understand (*i.e.,* not meaningful); thus, reducing these attribute features into being merely traditional low-level image/video feature descriptors. Furthermore, the discriminative power property and meaningful property are not independent. Some discriminative attributes could be meaningful and vice-versa. For example, the attribute "has hooves" is discriminative for distinguishing between dogs and sheep but the attribute "has four legs" is not. On the other hand, there is significant attention being given to the development of automatic attribute discovery approaches. These approaches focus on automatically discovering potentially meaningful attributes from data without the need for manual labelling. Although it is suggested that the discovered attributes are quite discriminative and meaningful, it is not entirely clear if they are truly meaningful. Quantitative and automatic evaluation methods to determine attribute meaningfulness are desirable in this case, since manual examination is both tedious and time-consuming. Research on this topic can start to shed light on how to automatically and effectively discover meaningful visual attributes without the huge cost of manual labelling.

The following research directions have not previously been extensively explored: (1) current works fail to address the problem of finding discriminative and meaningful attributes without involving human effort, and (2) there is no existing way to measure the meaning of discriminative attributes without involving human labelling effort. To that end, this research aims to devise visual attribute-based methods for traditional classification tasks and novel applications, including the attribute meaningfulness measurement, by dealing with these two shortcomings. In particular, this thesis addresses the following aspects: (1) automatic discovery of discriminative attributes from a set of meaningful attributes applied to zero-shot learning problems; (2) discovery of meaningful attributes by exploring ways to automatically quantify attribute meaningfulness; the proposed techniques are applied and tested on video keyword generation for video surveillance data, and (3) discovering meaningful and discriminative attributes in fully unsupervised scenarios via multi-graph clustering techniques. The main contribution of this thesis lies in the proposition of the method quantitatively evaluating the meaningfulness of automatic discovered attribute. This is inspired by our work on improving discriminative power of attribute by automatic image attribute selection. Then based on the guidances of the evaluating method, a novel category-level unsupervised attribute discovered method is proposed based on clustering framework. Moreover, application study has been performed in this research: an

automatically discovered attribute based keyword generation application is proposed in surveillance video scenarios.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

1. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell: **Automatic Image Attribute Selection for Zero-Shot Learning of Object Categories**: 22nd International Conference on Pattern Recognition 2014 (ICPR) 2014 [61].

2. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Kun Zhao, Brian C. Lovell: **Determining the best attributes for surveillance video keywords generation**: In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, IEEE, 2016 [63].

3. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell: **Automatic and quantitative evaluation of attribute discovery methods**: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) [62].

4. Liangchen Liu, Feiping Nie, Teng Zhang, Arnold Wiliem, Brian C. Lovell: **Unsupervised automatic attribute discovery method via multi-graph clustering**: 23nd International Conference on Pattern Recognition 2016 (ICPR) [60].

5. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell: **What Is the Best Way for Extracting Meaningful Attributes from Pictures?** Pattern Recognition (accepted 01.10.2016).

Publications included in this thesis

1. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell: **Automatic Image Attribute Selection for Zero-Shot Learning of Object Categories**: 22nd International Conference on Pattern Recognition 2014 (ICPR) 2014 [61]. Incorporated within Chapters 1–4.

Contributor	Statement of contribution
Liangchen Liu (Candidate)	Conceptual framework and structure (70%) Design of experiments (70%) Paper writing and editing (65%)
Arnold Wiliem	Conception and design of algorithm (15%) Design of experiments (15%) Paper writing and editing (15%)
Shaokang Chen	Conception and design of algorithm (10%) Design of experiments (10%) Paper writing and editing (10%)
Brian C. Lovell	Conception and design of algorithm (5%) Design of experiments (5%) Paper writing and editing (10%)

2. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Kun Zhao, Brian C. Lovell: **Determining the best attributes for surveillance video keywords generation**: In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, IEEE, 2016 [63]. Incorporated within Chapters 1–3, and 6.

Contributor	Statement of contribution
Liangchen Liu (Candidate)	Conceptual framework and structure (70%) Design of experiments (70%) Paper writing and editing (65%)
Arnold Wiliem	Conceptual framework and structure (15%) Design of experiments (10%) Paper writing and editing (15%)
Shaokang Chen	Conceptual framework and structure (10%) Design of experiments (10%) Paper writing and editing (10%)
Kun Zhao	Conceptual framework and structure (0%) Design of experiments (5%) Paper writing and editing (5%)
Brian C. Lovell	Conceptual framework and structure (5%) Design of experiments (5%) Paper writing and editing (5%)

3. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell: **Automatic and quantitative evaluation of attribute discovery methods**: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) [62]. Incorporated within Chapters 1–3, and 5.

Contributor	Statement of contribution
Liangchen Liu (Candidate)	Conceptual framework and structure (75%) Design of experiments (75%) Paper writing and editing (65%)
Arnold Wiliem	Conception and design of algorithm (10%) Design of experiments (10%) Paper writing and editing (15%)
Shaokang Chen	Conception and design of algorithm (10%) Design of experiments (10%) Paper writing and editing (10%)
Brian C. Lovell	Conception and design of algorithm (5%) Design of experiments (5%) Paper writing and editing (10%)

4. Liangchen Liu, Feiping Nie, Teng Zhang, Arnold Wiliem, Brian C. Lovell: **Unsupervised automatic attribute discovery method via multi-graph clustering**: 23nd International Conference on Pattern Recognition 2016 (ICPR) [60]. Incorporated within Chapters 1–3, and 7.

Contributor	Statement of contribution
Liangchen Liu (Candidate)	Conceptual framework and structure (70%) Design of experiments (75%) Paper writing and editing (65%)
Feiping Nie	Conception and design of algorithm (15%) Design of experiments (10%) Paper writing and editing (5%)
Teng Zhang	Conception and design of algorithm (0%) Design of experiments (0%) Paper writing and editing (5%)
Arnold Wiliem	Conception and design of algorithm (15%) Design of experiments (10%) Paper writing and editing (20%)
Brian C. Lovell	Conception and design of algorithm (5%) Design of experiments (5%) Paper writing and editing (5%)

5. Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell: **What Is the Best Way for Extracting Meaningful Attributes from Pictures?**: Pattern Recognition (accepted 01.10.2016). Incorporated within Chapters 1–3, and 5.

Contributor	Statement of contribution
Liangchen Liu (Candidate)	Conceptual framework and structure (70%) Design of experiments (70%) Paper writing and editing (65%)
Arnold Wiliem	Conception and design of algorithm (15%) Design of experiments (15%) Paper writing and editing (15%)
Shaokang Chen	Conception and design of algorithm (10%) Design of experiments (10%) Paper writing and editing (10%)
Brian C. Lovell	Conception and design of algorithm (5%) Design of experiments (5%) Paper writing and editing (10%)

Contributions by others to the thesis

The work contained in this thesis was carried out by the author under the guidance and supervision of his advisors, Professor Brian C. Lovell and Dr. Arnold Wiliem. Part of the work contained in this thesis was carried out by the author under the collaboration and discussions with Dr. Shaokang Chen and Professor Feiping Nie.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

A completion of any great work can never be made by anyone alone. It is significant to require the contributions from many different people. They devote a lot of time and effort, in their own ways, to help make this research possible. Their generosity professionalism kindness and inspiring discussion deeply affect and touch me. Here, I would like to express my most sincere appreciation to them especially the following:

My primary advisor, Professor Brian C. Lovell who enrolled me as the student for making this research possible. I appreciate your great support, advice and enlightening guidance throughout the whole research project, especially your great effort on proofreading my drafts. Your professional academic writing guidance helps me improve a lot on my scientific paper composition. Without his guidance, I cannot image how I can put the topic together. Thanks, Brian.

My co-advisor, Dr. Arnold Wiliem, for the great support on the problem discussion, experiment design, results analysis and paper writing. In total, you are such a good advisor who spends a lot of effort and even pays a great attention, to very detailed parts, to teach me how to manage the scientific research pipeline. Moreover, your incessant persistence sometimes may make me feel bad first but, eventually, I started to learn by myself and got some real knowledge.

My co-author, Dr. Shaokang Chen, for the inspiring discussion on the idea and experiment design, as well as your great effort on the draft proof-reading. Your rich experiences and thoughts on the background and theories of the research deeply inspired me. Without your sincere help, this research could not be running smoothly.

My former Ph.D. colleague Danny Smith. You are such a great predecessor for me, your great experiences and knowledge about the academic and engineering help me a lot. Thank you for not only forging the path ahead of me and show me the way but also affecting me with your personality of patience, optimism and dedication.

My former Ph.D. colleague Azadeh Alavi, Mahsa Baktashmotlagh and Sareh Shirazi, You have all done a great job on your Ph.D. research. Your achievements have become very good examples for the successor students like me.

My contemporary colleagues and co-authors, Teng Zhang, Kun Zhao and Johanna Carvajal Gonzalez who shared the hard but meaningful journey of the Ph.D. candidate with me for enjoyable coffee, dinner and lunch time, as well as meaningful discussion about the research topic of ours. I wish you all the best in your continuing research.

My new colleagues Allen Di, Wai San, Siqi Yang who brought the new blood stream into the group and make it forever vigorous. You are all very intelligent student and young scientists in the future. I learned a lot of new interesting knowledge from you from various backgrounds.

Last but not least, I would like to thank unconditional support and love from my parents, both financially and emotionally throughout my whole Ph.D. Candidate period. It is such an honour and fortune to be your son. Your patience, concerning and understanding not only made my way to my Ph.D. degree brighter but also constructed and polished my personality hence, made me an adequate student, a good man in society.

Keywords

Visual Attribute, Meaningfulness Metric, Attribute Discovering, Semantic Content, Scene Understanding.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080303, Computer System Security, 10%
ANZSRC code: 080109, Pattern Recognition and Data Mining 30%
ANZSRC code: 080104, Computer Vision, 60%

Fields of Research (FoR) Classification

FoR code: 0803, Computer Software, 10%
FoR code: 0801, Artificial Intelligence and Image Processing, 90%

Contents

List of Figures

List of Tables

List of Algorithms

Outline

Acronyms and Abbreviations

ACC	Accuracy
AMT	Amazon Mechanical Turk
ApAy	A-pascal A-yahoo dataset
ASUN	SUN Attribute dataset
AUC	Area Under the Curve
AwA	Animal with Attributes dataset
CCV	Columbia Customer Video
CRF	Conditional Random Field
DAP	Direct Attribute Prediction
DARL	Discriminative and Reliable Attribute Learning
DBC	Discriminative Binary Codes
GIST	A Holistic Representation of the Spatial Envelope
HIT	AMT Human Intelligent Task
HOG	Histogram of Oriented Gradient
IAP	Indirect Attributes Prediction
ITQ	ITerative Quantization
JND	Just Noticeable Difference
knn	*k*-nearest neighbor
KSH	Kernel-based Supervised Hashing
LBP	Local Binary Pattern
LDA	Latent Dirichlet Allocation

LSH	Locality Sensitivity Hashing
MAP	Maximum A Posteriori probability
MFCC	Mel-scale Frequency Cepstral Coefficient
MGCAD	Multi-Graph Clustering Attribute Discovery model
MPPCA	Mixtures of Probabilistic Principal Component Analysers
NMI	Normalised Mutual Information
PCA	Principal Component Analysis
pLSA	probabilistic Latent Semantic Analysis
RGB	Red-Green-Blue colour model
rgSIFT	*rg* Scale Invariant Feature Transform
RKHS	Kernel Hilbert Space
ROC	Receiver Operating Characteristic
SIFT	Scale Invariant Feature Transform
SOINN	Self-Orgnizing and Incremental Neural Networks
SPH	SPectral Hashing
STIP	Space-Time Interest Points
SURF	Speeded Up Robust Features
SVD	Singular Value Decomposition
USAA	Unstructured Social Activity Attribute dataset
UTTower	UT Tower aerial view dataset

Chapter 1

Introduction

> Do the difficult things while they are easy and do the great things while they are small. A journey of a thousand miles must begin with a single step.
>
> Lao Tzu

Chapter Summary: *Current high-level semantic features have drawn a large amount of attention from society since they are considered a possible way to provide the interface between human and machine. Among them, visual attributes are one of the most promising methods. In this research, we focus on visual attributes and their important properties, as well as propose several visual attribute-based applications, by answering three crucial research questions.*

Human language is the predominant means of communication. The utilisation of human language models in computer vision problems such as scene understanding [80], image or video description [58] could help in transferring knowledge acquired by human experts for solving these problems. People often say, "A picture is worth a thousand words." Turning this around, we could also say, "a thousand words are required to explain a picture." This latter form can be quite powerful to address many computer vision problems [28, 49, 29]. However, which words are the most expressive and discriminative to characterise the picture or the objects in the picture? This is not a hard problem for humans, but it is still an open question in computer vision research.

For instance, the active learning framework proposed by Kovashka *et al.* in [46] employs human knowledge to learn better models. Among these problems, visual attribute related tasks [49, 69, 121, 19, 50] are particular topics gaining interest in the research community as well as industry. Visual attributes have been proved to be promising because of their advantages *i.e.,* human understandable, machine computable and shared across classes.

Generally, visual attributes can be described as inherent properties/characteristics of images or visual information. Figure 1.1 illustrates an example of visual attributes in a blue car image. In the

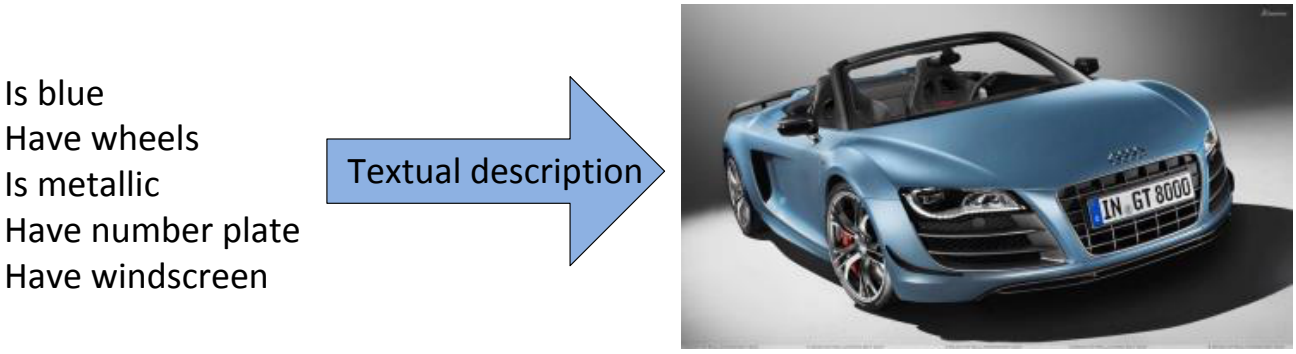


Figure 1.1: Illustration of a car image and its attributes

figure, this image contains five attributes: *is blue, have wheels, is metallic, have number plate, have windscreen*. These attributes can be used to describe the image. Furthermore, one can construct an image descriptor based on these attributes. Each element in the image descriptor, technically, defines the existence/absence of an attribute.

1.1 Goals and Challenges

The original goal of this thesis was to investigate the basic nature of visual attributes and their related novel applications.

Visual attributes are extremely useful as they are: (1) human understandable; (2) machine computable; and (3) shared across classes.

'Human understandable' means that they have meaningful names associated with them. Here we should note that the terms meaningful or meaningfulness are both terms in computer vision society and they are not related to the terms in database research. For example in database research, meaningful means that the changes between two trees in the database are in a succinct and descriptive way so the edit script of the changes are with high quality [15]. In the other words, the operations of changes can be more easily explained and interpreted if they are detected in a meaningful way. Meaningful in database research is also related to the recommendation system. There is an example [22] from meaningful ratings interpretation. The meaningful ratings interpretation can give a meaningful interpretation of ratings associated with the input items such as given a movie, it can output a set of meaningful factoids (for example, "male reviewers under 30 from NYC love this movie") rather than an overall rating aggregation. These are very different concepts from the "meaningfulness" or "meaningful" in this thesis.

Traditional images or visual descriptors and features do not have direct semantic meaning that is understandable by humans. This makes them almost impossible for humans to relate the features to some words. On the other hand, visual attributes are meaningful, and therefore human knowledge

could be efficiently transferred to solve challenging problems such as zero-shot learning [50] and active learning [6].

'Machine computable' indicates that the visual attributes can be detected and computed by a machine. There are several ways to construct the attributes and make them machine computable such as considering them as binary codes [49] or modelling them via ranking functions [78]. For example, in the binary case, attribute detectors are essentially concept detectors which detect the presence/absence of the specific concepts in an image or video. For these reasons, recently there are many attribute discovery methods that have been proposed to extract visual/image attributes [4, 86, 45].

'Shared across classes' means the categories are often related to each other so they process common visual attributes. For example, dogs and cats are two different animals, but they share common attributes such as *is a four-legged animal* and *has fur*. This means we can *extract* information about the common attributes from one class and map them into other disjoint classes [50].

Despite their advantages, visual attributes still suffer from a number of shortcomings: (1) the need for extensive labelling to train the attribute detectors; and (2) the lack of discriminative power. First, visual attributes [50, 80] could be extracted using binary classifiers acting as concept detectors, so they need extensive human labelling effort for creating training data to train these classifiers/detectors. This is because in traditional multi-class classification tasks, each sample only needs to be labelled once (*i.e.,* the label indicating to which category an image belongs), however, the labelling work for attributes requires each image to be labelled multiple times for each attribute in addition to its category label. It is extremely tedious, time-consuming, and expensive (*e.g.,* if there are 64 attributes, then each image should have at least 64 labels with one label for each attribute). Furthermore, in some specialised domains such as *Ornithology* [114], *Entomology* [109] and cell pathology [115], the human labelling task could be immensely expensive as only a few highly trained experts could conduct such work. Therefore, this drawback will constrain large-scale application or use in some special areas such as medical information processing or security where such effort is too expensive or intractable. Secondly, the image and visual attributes are often used as the features for the classification task [49, 69, 51]. Here, discriminative power is very important for the features because the good performance of classification tasks generally needs the input features to be as discriminative as possible. Discriminative power measures the separability of different sample points belonging to different classes in the feature space. Traditional attribute descriptors or attribute-based features do not often impose the discriminative power constraint for classification tasks [51, 81]. This would lead to poor classification performance.

As mentioned, one of the biggest challenges in using attribute descriptors is that enormous amount of training images with attribute labels are required to train attribute classifiers. To reduce the workload, automatic attribute discovery methods have been developed [4, 86, 95, 115, 123]. The primary aim of these works is to learn a function that maps the original image feature space into a binary

code space wherein each individual bit represents the presence/absence of a visual attribute. These attribute discovery methods are also closely related to hashing methods [34, 54, 113]. The difference is that, unlike automatic attribute discovery approaches, hashing methods focus more on how to reduce significantly the storage demand and computational complexity whilst maintaining system accuracy.

In computer vision research, determining expressive and discriminative words for picture and object description is still not appropriately solved. In spite of the merit of the new fashion of attribute discovery, there are still some questions left *i.e.,* are these automatically discovered attributes or binary codes meaningful? How can we evaluate an automatic attribute discovery method? By exploring these questions, we can begin to glean some insights on mechanisms required to extract meaningful attributes/binary codes.

Consequently, the research questions in this thesis focus on the following:

Q1: Can we automatically discover discriminative visual attributes that are still meaningful from images and video data?

Q2: Can we automatically discover visual attributes that are meaningful from a set of videos/images with category labels? If yes, how can the attribute meaningfulness be measured?

Q3: Can we automatically discover visual attributes from a set of videos/images without any supervision?

This research focusses on the feasible solutions and discussions on meaningfulness and discriminative power of visual attributes. Moreover, the theoretical background from the unsupervised scenarios for the novel automatic attribute discovery method is investigated.

1.2 Contribution

The works in this thesis first start from the discriminative power of visual attributes. Not all visual attributes are discriminative enough to distinguish between classes. For example, the attribute "fluffy" is not able to distinguish between dogs and cats. Therefore, a Discriminative and Reliable Attribute Learning (DRAL) method is proposed to select discriminative, yet meaningfully, reliable attributes from the set of meaningful attributes.

The discriminative power is not the only component that contributes to the attribute application. Meaningfulness of an attribute also plays an important role. For a traditional supervised attribute discovery method, the meaningful property is reflected by attribute prediction reliability, in other words, the generalisation error of the predefined attribute detectors. Recently, many new automatic attribute discovery methods [34, 54, 113] are proposed. The primary aim of these works is to learn a function that maps the original image feature space into a binary code space wherein each individual

bit represents the presence/absence of a visual attribute. In contrast to the human predefined attributes, the meaningfulness property is essential in this case because when the mapping function is designed just to increase the discriminative power in the output feature space, the individual binary bits (*i.e.,* one binary bit for one attribute) may not have any relationship with any semantic meaning; hence not meaningful. In order to study this further, extensive human effort is required to evaluate the meaningfulness of each binary bit or each attribute. Due to the tedious and time-consuming nature of traditional human involved evaluation methods, we propose a novel automatic meaningfulness metric that will become one of the *yardsticks* to measure attribute meaningfulness based on *shared structure* assumption [76, 77].

With the help of meaningfulness metric, in this thesis we also select the method that can automatically discover the most meaningful attributes. We apply this method to study the keyword generation problems in surveillance video scenarios and obtain better results over some other attribute discovery methods.

Finally, inspired by the insights from designing the metric, we propose a novel category-level unsupervised attribute discovery method utilising a multi-graph approach that preserves both local neighbourhood structure, as well as class separability.

Thesis contributions can be listed as follows:

1. A discriminative and reliable attribute learning method on visual attribute selection for better attribute-based classification performance and efficiency.

2. A novel automatic meaningfulness metric to measure attribute meaningfulness based on *shared structure*.

3. A keyword generation application in surveillance video scenarios based on the automatic attribute discovery method chosen by the attribute meaningfulness measurement.

4. A novel category-level unsupervised attribute discovery method via clustering framework to automatically discover visual attributes.

1.3 Thesis Outline

The section provides an outline of the entire thesis. For each chapter, there is a short introduction part and summary to guide the readers. The summaries outline the corresponding parts of the thesis. They provide the main ideas of this research. The collection of summaries is located on Page xxii.

The thesis includes the next 7 chapters. Figure 1.2 is a structure diagram that describes how different topics link together.

Chapter 2 introduces the background material of why visual attribute is important. In this part, the visual attribute will be introduced. In addition, its relationship to low-level features is discussed.

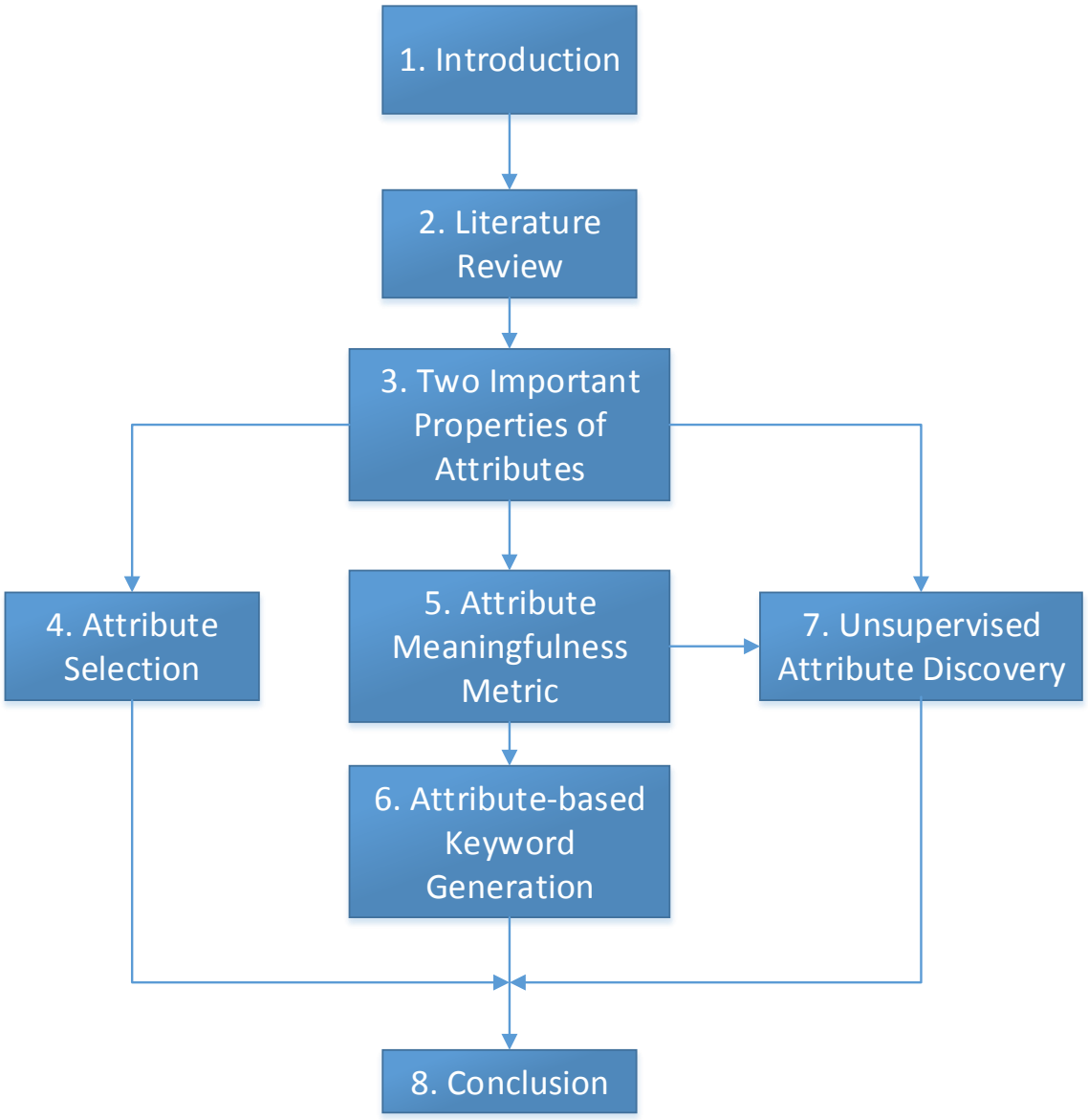


Figure 1.2: Flow chart illustrating the connection between the chapters in this thesis.

Moreover, different computational models of visual attributes and the various typical attribute related applications will be introduced in this part.

Chapter 3 will mainly focus on the specific problems that are solved in this research. The internal relationships of these problems are discussed based on two intrinsic natures of attributes *i.e.,* discriminative power and meaningful property. The methods proposed in this thesis are briefly introduced in this part to outline the highlights.

Chapter 4 introduces an automatic image attribute selection algorithm to improve the discriminative power of visual attribute. This algorithm is called Discriminative and Reliable Attribute Learning (DRAL). The main goal is to improve the performance of zero-shot learning [51] of object categories.

The method mainly considers and incorporates the two intrinsic natures of attributes *i.e.,* discriminative power and meaningful property for the human predefined visual attribute.

Chapter 5 investigates a more fundamental problem of how to measure attributes discovered automatically. With this metric, automatic quantitative evaluation can be performed on the attribute sets; reducing the enormous effort to perform the manual evaluation. In our evaluation, we gleaned several insights that could be beneficial in developing new automatic attribute discovery methods.

Chapter 6 introduces an automatic keyword generation method for surveillance video based on the automatic attribute discovery method. In this part, with the help of the method proposed in Chapter 5, we are able to find the automatic attribute discovery method that can discover the most meaningful attributes. Thus, the best discovery approach is employed to generate keywords for surveillance video scenario. This chapter shows it is possible to massively reduce the amount of manual work in generating video keywords.

Chapter 7 introduces a novel class-level unsupervised automatic attribute discovery method via multi-graph clustering. The novel method discovers more meaningful automatic attributes. From the insights of previous works, preserving local neighbourhood in the attribute binary space and increasing attribute binary feature discrimination will improve the meaningfulness of the attribute. In this method, the local neighbourhood structure is preserved by considering multiple similarity graphs; the class separability is achieved by incorporating the traditional clustering objective. The evaluation shows that the proposed method is able to discover more meaningful attributes than other unsupervised methods. Moreover, it is able to improve the performance in the clustering task.

Chapter 8 concludes this thesis with a list of the main contributions of the research, and outlines future directions for extension of the thesis.

Chapter 2

Literature Review

Yesterday is but today's memory,
and tomorrow is today's dream.

Khalil Gibran

Chapter Summary: *Visual attributes are novel high-level features, which currently have two models* i.e., *the binary attribute model and the relative attribute model. Meaningfulness and discriminative power are two important properties of visual attributes. A lot of works has been performed on each of them. However, there are few works that fully consider them together. Many attribute-based frameworks have been proposed in the research society for various applications, indicating the high-level feature is a promising research topic.*

2.1 Introduction

Following the previous introduction of the visual attributes and problems that we will address in this research, here we review the related literature to inspect the drawback of the current visual attribute works and collect the useful materials from related research.

Feature extraction plays a major role in the image and visual classification system pipeline. It serves as a transformation function mapping the images from their original high dimensional space into another space where the classification problem could be more easily solved. From this viewpoint, the visual attribute generally can be considered as a high-level feature or intermediate representation [98, 48, 28]. For example, in the attribute binary model [98, 31], the visual attributes are actually outputs of a series of attribute detectors/classifiers. Attribute detectors serve as the transformation functions mapping low-level features or descriptors into the attribute spaces. The result of the attribute extraction process is that each individual descriptor element has semantic meaning (*i.e.,* value 1/0 indicates presence/absence of an attribute concept).

Table 2.1: Works related to attribute models

Reference	Venue	Year	Overview
Ferrari *et al.* [31]	NIPS	2007	Binary or confidence-based attributes
Kmar *et al.* [49]	ICCV	2009	
Lampert *et al.* [50]	CVPR	2009	
Parikh and Grauman [78]	ICCV	2011	Relative attributes
Xiao and Jae Lee [118]	ICCV	2016	

Table 2.2: The works related to semantic attributes (CVPR: IEEE International Conference on Computer Vision and Pattern Recognition, ICCV: IEEE International Conference on Computer Vision, ECCV: European Conference on Computer Vision, NIPS: Annual Conference on Neural Information Processing Systems, WACV: IEEE Winter Conference on Applications of Computer Vision, TPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence, BMVC: British Machine Vision Conference, IJCV: International Journal of Computer Vision)

Reference	Venue	Year	Overview
Farhadi *et al.* [28]	CVPR	2009	Describing object face and clothes
Kumar *et al.* [49]	ICCV	2009	
Farhadi *et al.* [27]	CVPR	2010	
Chen *et al.* [17]	ECCV	2012	
Dhar *et al.* [24]	CVPR	2011	Attributes to predict aesthetics quality
Isola *et al.* [41]	CVPR	2011	Attributes for memorability
Isola *et al.* [40]	NIPS	2011	
Wang and Mori [112]	ECCV	2010	Relationship between objects and attributes
Kovashka *et al.* [47]	ICCV	2011	
Hwang *et al.* [39]	CVPR	2011	
Liu *et al.* [57]	CVPR	2011	Attributes for video representation
Rohrbach *et al.* [89]	ECCV	2012	
Ma *et al.* [69]	CVPR	2013	
Liu *et al.* [58]	WACV	2013	
Li *et al.* [56]	CVPR	2013	
Lampert *et al.* [50]	CVPR	2009	Zero-shot and one-shot learning
Yu and Aloimonos [124]	ECCV	2010	
Lampert *et al.* [51]	PAMI	2013	
Layne *et al.* [52]	BMVC	2012	Person re-identification
Kovashka and Grauman [45]	IJCV	2015	Crowdsourcing
Deng *et al.* [23]	TPAMI	2016	
Patterson *et al.* [81]	IJCV	2014	Dataset construction

Table 2.3: The works related to discriminative attributes (BTAS: International Conference on Biometrics Theory, Applications and Systems)

Reference	Venue	Year	Overview
Farhadi *et al.* [28]	CVPR	2009	Describing object
Wang and Mori [112]	ECCV	2010	Relationship between objects and attributes
Kovashka *et al.* [47]	ICCV	2011	
Hwang *et al.* [39]	CVPR	2011	
Liu *et al.* [57]	CVPR	2011	Attributes for video representation
Rohrbach *et al.* [89]	ECCV	2012	
Ma *et al.* [69]	CVPR	2013	
Li *et al.* [56]	CVPR	2013	
Su *et al.* [98]	ECCV	2012	Large-scale image classification
Fu *et al.* [32]	PAMI	2013	Sparse and incomplete attribute labels
Yu *et al.* [123]	CVPR	2013	Automatically designing discriminative category-level attributes
Wang *et al.* [110]	TIP	2016	Attribute specific feature selection
Yu and Grauman [122]	ICCV	2015	Fine-grained visual search
Samangouei *et al.* [92]	BTAS	2015	User Authentication

The brief summaries of current works in attribute research and the applications are presented in Table 2.1, Table 2.2, and Table 2.3.

Table 2.1 presents two typical models of visual attributes: (1) the traditional binary or confidence-based attribute model; and (2) the relative attribute model. The traditional binary or confidence-based attributes are in the form of binary codes or real numbers ranging from -1 to +1. They represent the presence/absence of the attributes or the strength of the presence of attributes. The relative attributes are constructed through learning a ranking function for each attribute. The ranking functions then predict the relative strength of the attributes.

Then, since it has been mentioned in Section 1.1 that the discriminative power and semantic meaning are two significant properties of the image and visual attributes, we group the current works into the two tables: Table 2.2 and Table 2.3. Table 2.2 presents the works that train attribute classifiers using manually labelled images. Generally, these works do not aim to improve attribute discriminative power. Table 2.3 summarises the works which aim to discover discriminative attributes for a particular application domain. Their main applications are briefly introduced in the overview column. Unfortunately, despite their success in finding discriminative attributes, there is no constraint imposed to guarantee that the attributes are meaningful. More discussion about the details will be elaborated in the next several sections.

In the light of these facts, it is safe to say that the studies in devising methods that learn discriminative and semantically meaningful attributes are still largely unexplored. One of the biggest advantages of having both properties is that one can use the feature to improve the system performance as well

as infuse human knowledge into the system. In addition, a textual description, which is beneficial for some computer vision problems such as in advanced surveillance systems, could be easily produced.

In the following subsections, we first briefly review the research works of low-level features as they are the foundation of visual attributes. Next, the current two types of attribute modelling in Table 2.1 are discussed. After that, the representative works in Table 2.2 and Table 2.3 are elaborated. Then, towards the research question on automatic attribute meaningfulness evaluation, we start to gather some useful materials and inspirations related to our work. Finally, we refer to some application examples of attribute-based frameworks.

2.2 Low-level Features

Since the visual attribute generally can be considered as a type of semantic high-level feature or intermediate representation, here we discuss some works on low-level features to distinguish between visual attributes and low-level features. The low-level feature has been widely explored and has been used in various applications [102, 75, 21]. Vast existing works aim to develop effective methods extracting features from raw pixel-level basic visual information such as intensity and colour. For instance, the Scale Invariant Feature Transform (SIFT) [67] aims to extract features possessing invariant properties such as location, scale, rotation, and affine transformations. In addition, *rg* Scale Invariant Feature Transform (rgSIFT) [102] is an improved feature extraction method based on SIFT. It adds the *r* (red) and *g* (green) chromacity components of the normalised RGB colour model, which is already scale-invariant. Another notable example is the Histogram of Oriented Gradient (HOG) [21] which is a feature descriptor counting occurrences of specific gradient orientations in localised portions of an image. Local Binary Pattern (LBP) [75] is also a good example, which presents a theoretically very simple, yet efficient, multi-resolution approach to gray-scale and rotation invariant texture classification based on local binary patterns, nonparametric discrimination of sample, and prototype distributions. Speeded Up Robust Features (SURF) [3] is also a prevalent and effective low-level feature extraction method which is a novel scale- and rotation-invariant interest point detector and descriptor. It is achieved by relying on integral images for image convolutions by building on the strengths of the leading existing detectors and descriptors and by simplifying these methods to the essential. Despite the excellent performance reported for various vision applications, these feature descriptors generally do not have direct semantic meaning. In other words, different from visual attribute, these descriptors are difficult to interpret by human.

2.3 Attribute Models

In this research, we mainly focus on the binary attribute model. To be comprehensive, here we introduce the computational models of attributes. There are two main frameworks contributing to the computational models of visual attributes.

One of the frameworks considers each attribute as a binary code, which represents the presence/absence of a specific attribute, or a real number range from -1 to 1 which represents the strength of the presence of the attribute [49, 50, 31]. For example, in this model, suppose there is a set of attributes: $A = \{a_m\}, m \in \{1, ..., M\}$ and image features $\{x_i\}, i \in \{1, ..., N\}$, the corresponding attribute detectors f_m of A can be trained from the image feature x_i from the i-th image and its attribute labels of this set for those images. Then the detectors f_m can be used to extract visual attribute for the arbitrary image as:

$$x_i \xrightarrow{f_m} \{+1, -1\} \tag{2.1}$$

where +1 and -1 respectively represent the presence and absence of the m-th attribute.

The other framework is the relative attribute model [78], which learns a ranking function per attribute, given training data stating how object/scene categories relate according to various attributes. The learned function then predicts the relative strength of each property/attribute in novel images. More specifically, this model considers the additional detailed supervised condition which is the order and relative relation (*i.e.,* similar or not) between the samples for an attribute besides the presence or absence. For each attribute a_m the supervision is as $O_m : w_m(x_i) \succ w_m(x_j)$ and $S_m : w_m(x_i) \sim w_m(x_j)$. Here O_m indicates the set of samples in which the m-th attribute of sample x_i is stronger than x_j, and S_m indicates the set of samples in which the sample x_i and x_j have similar relative strength of the m-th attribute. The scoring function w_m can be trained to extract the m-th relative attribute as:

$$x_i \xrightarrow{w_m} \text{Real value (strength of attribute presence)} \tag{2.2}$$

For example, in Figure 2.1, the scene in (a) is more "natural" than that in (b), Scarlett Johansson in (c) smiles less than Miley Cyrus in (d).

Finally, these relative attributes can be jointly used in zero-shot learning or image description tasks. Generally, the binary code model is still the mainstream model for the majority of current works, for this model is simple to use in applications and easy to be interpreted by humans.

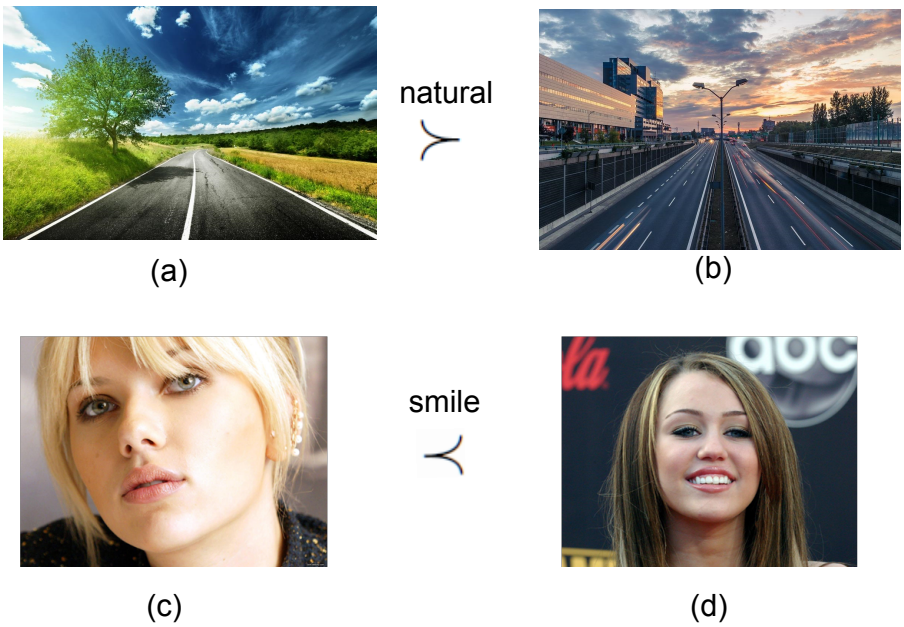


Figure 2.1: Illustration of the relative attributes, the first row describes the attribute "natural" for scene images, the second row presents the attribute "smile" for face images.

2.4 Representative Works of Meaningful and Discriminative Attribute

Just at a glance in Table 2.2 and Table 2.3, it is not hard to find out that there are some shortcomings of current works. We can see that many works either consider the meaningful property or consider the exclusively discriminative power without explicit semantic meaning.

For example, Yu *et al.* designed a category-level discriminative attribute learning algorithm according to category-separability and learnability [123]. However, their method cannot be used to describe images with concise semantic meaning. There is further effort to build a new category-attribute matrix for the zero-shot learning task. Kankuekul *et al.* propose an online incremental attribute-based zero-shot learning method using Self-Organising and Incremental Neural Networks (SOINN) as the learning mechanism [43] and dramatically reduces the computation time. Yet this method needs a great deal of human effort and fails to consider the inherent discriminative power of existing attributes. Lampert *et al.* propose two general frameworks of attributes-based zero-shot learning, Direct Attributes Prediction (DAP) and Indirect Attributes Prediction (IAP), which serve as a foundation for the existing works [50, 51]. However, the discriminative power of attributes has not yet been considered. Kumar *et al.* propose two methods for face verification [49] however, they fail to discover the discriminative power of these special facial attributes. Chen *et al.* [17] propose a practical method to describe clothing using semantic attributes. Yet, they are not explicitly maximising the attribute discriminatory power. Liu *et al.* introduce concept attributes into video event recognition for the first

time [58], however discriminative power has not yet been considered in that work. Farhadi *et al.* propose to use random comparisons and within category prediction to learn the discriminative and semantic attributes respectively [28]. Nonetheless, the resulting set of discriminative attributes is not guaranteed to have semantic meaning; defeating one of the primary purposes of using attribute descriptors. Recently, deep learning models have attracted a lot of attentions form the society, There are also some attribute related works [125, 20, 66] based on deep framework, which have gained stronger reliability. Thus the strong attribute reliability is very likely to benefit the discriminative power for applications such as classifications. For example, Zhang *et al.* [125] proposed a new method which combines part-based models and deep learning by training pose-normalized Convolutional Neural Networks (CNNs). This method has shown substantial improvement over the state-of-art methods on challenge attribute classification task. Chuang *et al.* [20] proposed a deep attribute network (DAN) to address the issue of representing input image on the basis of high-level features (attributes) that carry semantic meaning which humans can understand. The model can outputs the attributes of a given input image without performing classification. Liu *et al.* [66] proposed a novel deep learning framework for attribute prediction in wild. This framework cascades two CNNs, LNet and ANet, which are finetuned jointly with attribute tags, but pre-trained differently. This framework not only outperforms the state-of-the-art with a large margin, but also reveals valuable facts on learning face representation. However, they still have not tried to model the attribute meaningfulness in mathematical way which is one of the main contributions of this research. Furthermore our proposed methods can be easily extended using the CNN framework such as using the CNN features to construct the meaningful attributes subspace.

Only a few works combine these two parts together. Some researchers [112, 47, 39] study the relationship between object recognition and attributes learning, however, they need human efforts in the loop. Recently there are growing works [69, 56, 57, 89] on attribute-based video applications, however, those works restrict their scope to a very specific application area and need intensive human labelling efforts involved beforehand.

2.5 Evaluation of Meaningfulness of Visual Attribute

As two important properties of visual attributes, there are perfect evaluation methods to measure discriminative power (the performance of classification task of categories). However as to meaningfulness, as mentioned in the research question, how the attribute meaningfulness can be measured automatically by machine is still not properly addressed. Here we discuss some related works that will assist to pave the way to automatic evaluation of meaningfulness of visual attribute.

Evaluation of visual attribute meaningfulness is traditionally conducted by manually checking the presence/absence of consistent identifiable visual concepts in a set of given images. This task

usually requires a large-scale human labelling effort. A system such as the Amazon Mechanical Turk (AMT) [1] is able to handle this task for small datasets. However, since this process needs to be repeated whenever new attributes are discovered or novel methods are proposed, this manual process is ineffective and expensive. In our case, the AMT Human Intelligence Task (HIT) is to evaluate the meaningfulness of attributes by examining corresponding positive and negative images according to each attribute. The average time of each worker spent on this typical HIT is 2 minutes [80]. Then an AMT worker may require 320 minutes to evaluate 32 attributes discovered by 5 different methods (*i.e.,* $32 \times 5 \times 2 = 320$ minutes). The time spent could increase significantly if statistically reliable results are desired by increasing the number of AMT workers.

Unequivocally, it is more desirable to develop an automatic approach which is more cost-effective, and less labour intensive and time consuming, to evaluate the meaningfulness of the set of discovered attributes. The task of measuring the *attribute meaningfulness* of discovered attributes is similar to the task presented in the Turing Test [100]. In this task, we would like to measure how much a machine could provide responses like a human being. If a machine could respond like a human being, it means that the results produced must have significant meaning. Unfortunately, the Turing Test still requires a human judge who actively engages with the machine.

To that end, several works [2, 91, 100] aim to devise an automated Turing Test that follows the framework of this famous test but replaces the human judge by another machine. A notable example is CAPTCHA [106], which is very prevalent in web security applications. This technique lets a machine be the judge issuing the test to determine whether the subject is a human. Generally, CAPTCHA provides a challenge in the form of an image containing numbers or characters, which are difficult to be identified by current machines. The main assumption in CAPTCHA is that machine recognition will not be as good as human.

Our work on the measurement of meaningfulness of visual attribute can be interpreted as an instance of the automated Turing Test as follows. We are testing a set of automatic attribute discovery techniques by giving challenges in the form of images. These techniques are then giving us a set of attributes. We will automatically verify the meaningfulness through the positive and negative images generated from each attribute classifier. Note that if we have human observers performing the verification instead of machines, then this becomes an instance of the standard Turing Test. To perform the automated Turing Test, there has to be a measurement to determine which automatic attribute discovery technique is 'good' and which one is 'not that good'.

Some unsupervised semantic visual representation learning works [38, 36, 86, 115] have indicated that it is possible to discover the meaningful concepts without supervision from data itself with or without side information. Chen *et al.* [38] introduce a simple, yet powerful unsupervised approach to learn and predict visual attributes directly from data. With the help of deep Convolutional Neural

[1] www.mturk.com

Networks (CNNs), they train to output a set of discriminative, binary attributes often with semantic meanings. Hong *et al.* [36] propose a novel algorithm to cluster and annotate a set of input images with semantic concepts jointly. They employ non-negative matrix factorisation with sparsity and orthogonality constraints to learn the label-based representations with the side information (a labelled reference image database) obtaining promising results.

All of these works imply there may be some potential relations between meaningful concepts. Fortunately, the shared structure assumption among meaningful attributes proposed in [77] can serve as the foundation of the automatic measurement. Based on this assumption, Parikh and Grauman propose an active learning approach that uses Mixtures of Probabilistic Principal Component Analysers (MPPCA) [99] to predict how likely an attribute is nameable. Nevertheless, their work only focusses on deciding whether an attribute is nameable or not. Their work does not tackle the problem of quantitatively measuring the attribute meaningfulness. In addition, this approach requires human interaction to populate the nameability space. Thus, their method is not suitable for addressing our goal (*i.e.,* to automatically evaluate the meaningfulness of attribute sets).

In our work on attribute-based video keyword generation, the shared structure assumption is utilised. In particular, the work in Chapter 6 proposes a selection approach of attribute discovery methods to assist attribute-based keyword generation for video description from surveillance systems. However, the work does not consider quantitative analysis of the meaningfulness of the discovered attributes (*e.g.,* how much meaningful content is contained in a set of automatically discovered attributes). In addition, the characteristics of the meaningfulness of attributes may vary to some extent.

2.6 Application of Attribute-based Frameworks

The unique benefit of the visual attribute has brought about many successful attribute-based applications for various computer vision tasks such as face verification [49], Complex Event Detection [69], Human Action Recognition [121], visual knowledge extraction [19], and zero-shot learning [50]. For instance, Lampert *et al.* showed that the image attribute descriptor could be used to address the zero-shot learning [51, 50]. Here, a human can easily define the representation for each category in the test without any training image because each element of the descriptor has semantic meaning. Moreover, Parikh and Grauman proposed the notion of relative attributes such as *larger* or *more open space*, which could be understood as an adjective comparing two images [78].

Even from a more broad area, the visual attribute is still a powerful tool to provide a better computational model for the semantic or visual application. For example, Silberer *et al.* [97] show that visual attribute is beneficial for distributional models of lexical meaning together with the traditional text model. The bimodal models they proposed give a better fit to human word association data compared to amodal models and word representations based on handcrafted norming data.

Attribute-based models are also used in particular application scenarios such as medical image analysis and video surveillance. In medical image analysis, attributes can serve as the rationale understood by human experts for medical image classification. For example, Wiliem *et al.* proposed to use discriminative attributes for HEp-2 specimen image classification [115]. In video surveillance analysis, attributes can serve as the human-computer interface for anomaly detection or people searching in video surveillance. For example, Feris *et al.* proposed an attribute-based People Search system for video surveillance [30]. The semantic meaning of visual attribute can show the rationale for the classification, which may be detected by human experts. Yu and Grauman [122] proposed to develop a Bayesian local learning strategy to infer when images are indistinguishable for a given attribute. Borrowing the concept of the model of Just Noticeable Difference (JND) from psychophysics, this work aims to propose a solution to infer when two images are indistinguishable for a given attribute by introducing a Bayesian approach that relies on local statistics of orderability. To our knowledge, the framework offers the first attempt to unify a notion of equality into relative attributes during inference. Xiao and Jae Lee [118] developed a novel formulation that combines a detector with local smoothness to discover a set of coherent visual chains across the image collection. They use an efficient way to generate additional chains based on the initially discovered chains, determine the most relevant visual chains, and then create an ensemble image representation for the attribute model.

2.7 Related Works in Video Keyword Generation

In this thesis, we also apply the attribute model into a novel application scenario: surveillance video keyword generation. To the best of our knowledge, this is the first work employing the attribute model for video keyword generation in surveillance.

In this application scenario, there are only few methods proposed recently that deal with video keyword and description extraction [87, 90, 111, 104, 120]. For instance, Rohrbach *et al.* [90] proposed to generate a rich semantic representation of the visual content such as object and activity labels. They employed the Conditional Random Field (CRF) to model all the input visual components. In [87], they extended their work to a three-level-of-detail video description scheme. Then they applied a machine translation framework to generate the natural language using the semantic representation as sources. Unfortunately, this model cannot be used to address our problem due to the extensive manual labelling work required.

To that end, some researchers rely on hierarchical probabilistic models. Wang *et al.* [111] and Varadarajan *et al.* [104] employ Latent Dirichlet Allocation (LDA) and probabilistic Latent Semantic Analysis (pLSA) respectively to perform unsupervised activity analysis. However, these methods can only be applied to the bag-of-words framework. This means, more powerful features such as Fisher vectors [82] cannot be used directly.

Only one work specifically targets automatic video description problems in surveillance videos. Xu *et al.* [120] develop a novel distributed multiple-scene global understanding framework that clusters surveillance scenes by their ability to explain each other's behaviours. However, their work only focusses on the multiple-scene case and again, utilises hierarchical probabilistic models.

Chapter 3

Meaningful Property and Discriminative Power of Attributes

> To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science.
>
> Albert Einstein

Chapter Summary: *Meaningful property and Discriminative power of attributes are the main theme of this research. The works that are conducted in this thesis are all around the two important properties of visual attributes. In this part, we focus on the internal meaning of the two properties and describe the structure based on them that links different works and topics together.*

3.1 Introduction

Based on the foundation of visual attribute introduced in the previous chapters, we can start to put our emphasis on the specific problems. In this section, we get on to the details of properties of the visual attributes *i.e.,* semantic meaningful property and discriminative power property. We will also briefly introduce how we can employ these properties to discover better attributes as well as to devise the new attribute-based application and their relationship.

3.2 Meaningful Property of Visual Attribute

Meaningful property indicates how meaningful to human perception is the concept of the visual attribute of a group of images or videos. In the thesis, the term meaningfulness and semantics are

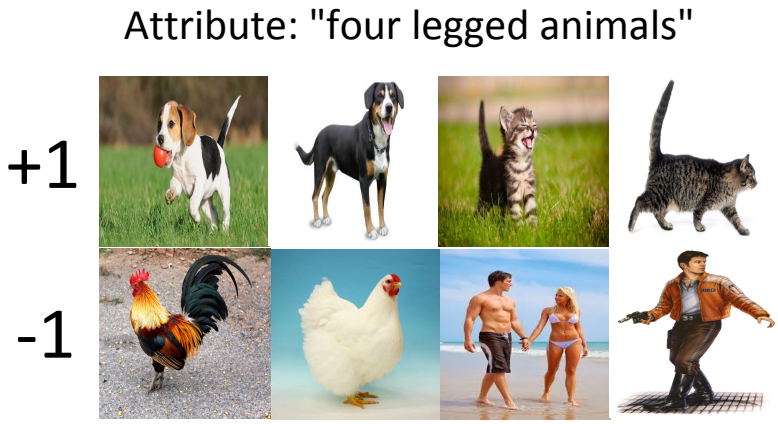


Figure 3.1: Illustration of meaningfulness property of visual attribute

interchangeable. In the real world situation, the differences between meaningfulness and semantics lie on the ability whether humans can verbally express the visual concept using human language. If yes, we consider the visual concept is visually semantic. However, if the visual concept can really be perceived but hard to explain using simple language, we still consider it as visually meaningful. We simplify the problem and consider them as one concept in the thesis. One important aspect in meaningful property is that the property can only be evaluated when the attribute extraction results of a set of images/videos are given. For instance, one may find an attribute to be meaningful when its value is +1 for images either containing cats or dogs and -1 for images containing chickens and humans. This may mean the attribute can be related to the attribute "four legged animals". Figure 3.1 illustrates this concept. Due to the various applications and attribute types, the meaningful property has a different description in different cases.

In the supervised case where manual attribute labels are provided, the attribute detectors are trained using these labels. Therefore, the meaningful property of attribute should be maximised. Admittedly, human perceptions of different annotators may have nuances (the extent of which could further alleviate the problem). However, manual attribute labels are still the ideal representations of meaningful property of the visual attribute. Since machines are less capable than humans to recognise the concept of the visual attribute, the meaningful property is jeopardised by the accuracy of the attribute classifier. Therefore, the meaningful property of the visual attributes in this scheme is determined by the accuracy of the attribute classifier (it can also be interpreted as the reliability of the visual attribute when employing it to novel testing samples, discussed in Chapter 4). In this thesis, this problem is investigated together with the discriminative power for automatic image attribute selection in zero-shot learning of object categories. Compared with the usage of the whole pool of meaningful attributes, the few visual attributes with the most discriminative and meaningful are automatically selected to improve the performance of zero-shot recognition of object categories. The results show the meaningful property is very useful when applied to the class recognition scenario. Details of this work is introduced in Chapter 4.

When there are no attribute labels provided, visual attributes can be discovered by various unsupervised methods, or methods with side information such as category labels. In this case, the meaningful property is more important, because the visual attributes are learned automatically from data structure with, or without, the prior knowledge from category labels. They may be used as very discriminative mid-level features for application. However, humans would not know if the discovered attribute (the split of the data) has obvious meaning before it is examined by a human himself or herself. How to evaluate the meaningfulness of this automatically discovered attribute is an important problem. By exploring this problem, we can begin to glean some insights on the mechanisms required to extract meaningful attributes. Furthermore, the traditional way of evaluating the meaningfulness of the visual attributes relies on manual examination *i.e.,* a human worker needs to examine the positive and negative samples of a specific attribute. This is apparently a tedious, time-consuming, and expensive evaluation method. It is desirable to devise an automatic evaluation method to alleviate the working load on human workers. In this research, we devise an attribute meaningfulness metric based on the *shared space assumption* between meaningful attributes for human perception to tackle this problem. The metric is mathematically formed by calculating the reconstruction error between the human predefined attribute and specific automatically discovered attributes. With the metric, quantitative evaluation of the meaningfulness of an automatically discovered attribute becomes possible. This work is elaborated in Chapter 5. For further validating the usefulness and the effectiveness of the proposed automatic attribute evaluation method, a novel application of keyword generation for surveillance video is proposed. The attribute meaningfulness evaluation method is applied to determine which automatic attribute discovery method can discover the most meaningful attributes set to generate meaningful output keywords for the videos. This work is introduced in Chapter 6.

Note that, since binary attributes are the most prevalent attribute model, we only consider this type of attribute. However, the semantic meaningful property can also be extended to the relative attribute model since human workers can also perceive an automatically discovered relative attribute through interpreting the samples displayed in a special way. For example, they are ordered in the direction perpendicular to a hyperplane that represents a specific attribute.

3.3 Discriminative Power of Visual Attribute

Different from the meaningful property of visual attribute, we consider the discriminative power as the ability of visual attribute feature to discriminate between the image classes. Discriminative power and meaningful property are not mutually exclusive to each other. They can benefit from each other, as some specific attributes can be perceived by human workers as the characteristics of a category used for distinguishing between the category and others, *e.g.,* shape of feet can be used to discriminate between dogs and deer, man-made objects can be used to discriminate between city view scenes and

natural outdoor scenes. This property is important for visual attributes as a powerful tool to use for some intelligent applications such as classification, image retrieval, and zero-shot learning. The discriminative power property has the consistent meaning for both supervised pre-trained attributes and automatically discovered attributes. Usually, automatically discovered attributes have better discriminative power since they are learned from the specific structure of the data. However, the pre-trained attributes, since they are directly learned from human knowledge, lack the ability to capture the intrinsic data structure, so are poor in discriminative power. To improve the discriminative power of pre-trained supervised attributes is a significant issue. As mentioned in the previous section, in this research, we target this problem by automatically selecting the most discriminative attributes in the pre-trained attribute pool to improve the performance of attribute-based classification.

On the other side, the automatically discovered attribute, even if it may be a more discriminative feature for some applications, the lack of meaningfulness to humans constrains their broader usage on other applications such as video or image description and active learning. To attack this problem, we propose in Chapter 7 a novel unsupervised attribute discovery method utilising a multi-graph approach that preserves both local neighbourhood structure as well as class separability. From the insights of our work in Chapter 5, (1) preserving local neighbourhood in the attribute binary space will increase the meaningfulness, and (2) increasing attribute binary feature discrimination will improve the meaningfulness. The results, using the proposed meaningfulness metric in Chapter 5, show the proposed method outperforms the other comparative unsupervised attribute discovery methods.

3.4 Summary

The discriminative power and meaningfulness property of visual attribute can construct a coordinate of two dimensions, illustrated in Figure 3.2. The more discriminative power the obtained attributes have, the more effectively the attributes can serve as features for a classification system (from a machine's viewpoint). Meanwhile, the more expressive the meaningfulness of the attributes is, the more understandable the attributes are for human (from human-being's viewpoint). In addition, the reliability is also a reflection of the meaningfulness property of pre-defined supervised attributes, which is also a significant prerequisite for decent performance of attribute-based systems.

The works we have completed in this thesis and other representative methods are summarised in Figure 3.2.

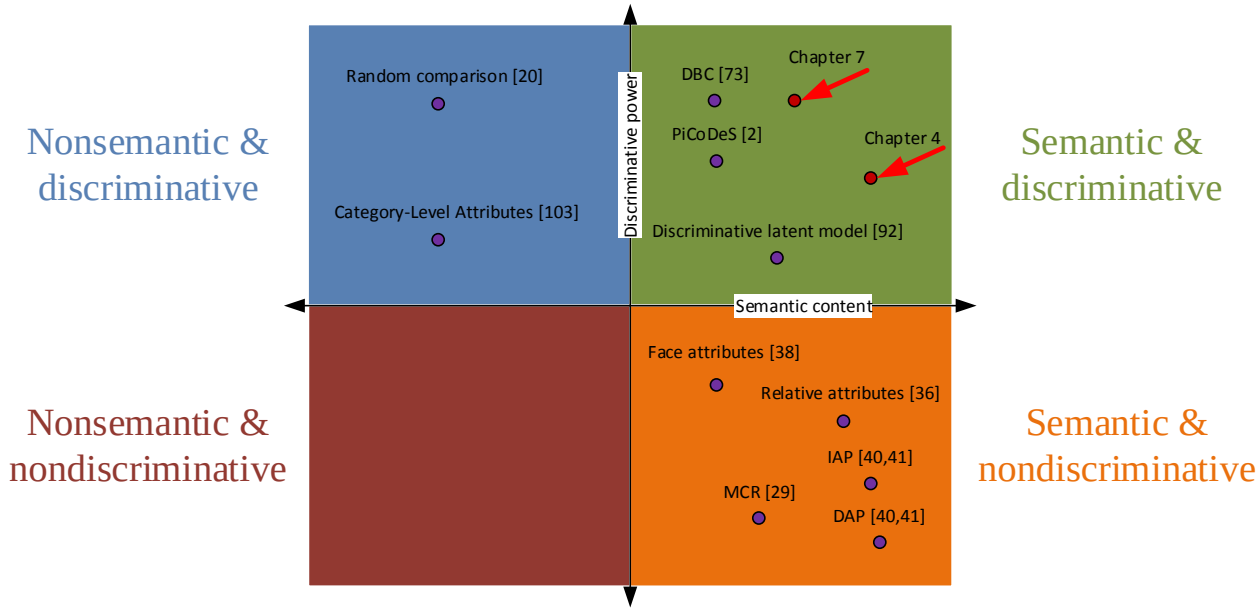


Figure 3.2: Illustration of two dimensions of the properties of visual attributes and the estimated position of our methods

Chapter 4

Improving Discriminative Power of Attribute by Automatic Image Attribute Selection

> ... though I have called this principle, by which each slight variation, if useful, is preserved, by the term o.
>
> Charles Darwin

Chapter Summary: *We propose a novel greedy algorithm called Discriminative and Reliable Attribute Learning (DRAL) which selects a subset of attributes from the pre-defined attribute pool. The subset maximises an objective function incorporating the two properties* i.e., *discriminative power and semantically meaningful property (expressed as the reliability in this case). We compare our proposed system to the recent state-of-the-art approach, called Direct Attribute Prediction (DAP), for the zero-shot learning task on the Animal with Attributes (AwA) dataset. The results show that our proposed approach can achieve similar performance to this state-of-the-art approach while using a significantly smaller number of attributes.*

4.1 Introduction

Following the roadmap guided by the meaningfulness property and discriminative power of visual attributes, in this chapter, we investigate the problem in the attribute-level supervised case. More specifically, we propose to search for the most meaningful (*i.e.,* reliable in the supervised case) and discriminative attributes as the high-level semantic feature for the classification problem. Through the investigation and analysis, we are able to answer the research question 1 in Section 1.1.

Feature extraction is one of the prominent tasks in the image classification system pipeline. It serves as a transformation function mapping the images from their original high dimensional space

to another space where the classification problem could be easier to solve. There are many works aimed to develop such a good transformation function. For instance, the Scale Invariant Feature Transform (SIFT) [68] aims to extract features possessing invariant properties such as location, scale, rotation, and affine transformations. Another notable example is the Histogram Oriented Gradient (HOG) [21] which is a feature descriptor counting occurrences of specific gradient orientations in localised portions of an image. Despite their excellent performance reported for various vision applications [101, 21], these feature descriptors are very difficult to be interpreted by humans. Although each of their elements may have a relationship such as the gradient magnitude, they do not have a direct relationship to the high-level semantic concepts related to the problem domain [55, 65].

Image attributes can be described as inherent properties/characteristics of an image. For instance, a car image could have the following attributes: *is blue*, *has wheels*, *is metallic*. In this case, one could represent an image with a set of image attributes present in an image. Technically, each element in the descriptor defines the existence/absence of a specific image attribute. It can be detected by attribute detectors tested on the low-level features mentioned above. Attribute detector is basically a binary classifier trained beforehand. As such, one needs to construct different training sets for each attribute detector, which could be expensive. To that end, one could use a crowdsourcing approach which could minimise the cost by using the Amazon Mechanical Turk (AMT) [1][77, 79]. Here, we can ask people on the internet to describe the images by words. Generally, the set of attributes found from this process is not necessarily discriminative for the vision task. This is due to the fact that it is difficult for humans to manually identify a set of discriminative attributes for a classification task which has a large number of categories.

The image attribute representation is successfully applied in various vision tasks such as face verification [49], complex event detection [69], human action recognition [121], visual knowledge extraction [19], and zero-shot learning [51]. Moreover, Parikh and Grauman proposed the notion of relative attributes such as *larger* or *more open space* which could be understood as an adjective comparing two images [78].

In this chapter, we focus on the attribute-based zero-shot learning problem. Zero-shot learning [51] is the problem of object recognition when the testing categories do not have any training examples. However, humans can easily define the attribute representation for each test category without any training image due to the fact that each element of the attribute descriptor has semantic meaning. For instance, Lampert *et al.* showed that the image attribute descriptor could be used to address the *zero-shot* learning [51]. They proposed two general frameworks of attributes-based zero-shot learning, Direct Attributes Prediction (DAP) and Indirect Attributes Prediction (IAP). However, in their work, attribute discriminative power has not yet been considered.

[1] www.mturk.com

Several works proposed approaches to automatically discover discriminative attributes [28, 86, 123]. These approaches are very similar to some feature selection works [12, 11]. Here, attribute detectors are jointly learned with the image classifier in the max-margin framework. For instance, Farhadi *et al.* proposed to use random comparisons and within category prediction to learn the discriminative attributes respectively [28]. Nonetheless, the resulting set of attributes is not guaranteed to have semantic meaning; defeating one of the primary purposes of using attribute descriptors. Yu *et al.* designed a category-level discriminative attribute learning algorithm according to category-separability and learnability [123]. However, their method also cannot be used to describe images with concise semantic meaning due to the design of category-attribute matrix. There is also significant human effort required to build a new category-attribute matrix.

Contributions: The aim of this chapter is to discover the set of semantic attributes which are also discriminative and reliable for the given classification task. To that end, we propose a discriminative selection algorithm which takes as input the image attributes discovered from the manual process via the AMT. There are two main advantages of using the proposed approach: (1) the feature dimensionality can be significantly reduced which simplifies the classification process and (2) the selected attributes can potentially improve the system performance due to the fact that the selection is based on the attribute discrimination power. The algorithm selects the subset of semantically meaningful attributes maximising two attribute properties: attribute discriminative power and attribute reliability. Attribute discriminative power is related to the property of the attribute descriptor to separate images of different categories. Therefore, we relate the attribute reliability to the error produced during the attribute descriptor extraction (training) process. We apply our method to the zero-shot learning problem investigated in [51]. We show that by applying our algorithm we can decrease the dimensionality of the attribute descriptor by 35%, achieving better performance than the state-of-the-art approach proposed in [51].

We continue this chapter as follows. Section 4.2 presents the proposed attribute properties. Then we describe the proposed algorithm in Section 4.3. The experiment and results are discussed in Section 4.4. Finally, the main findings and future direction are presented in Section 4.5.

4.2 Property of Image Attributes

Each element of an image attribute descriptor defines the existence/absence of an image property [51]. Generally, in an image classification task, each image is represented by the same set of image attributes. Let $\mathbf{z_i} \in \{0,1\}^B$ be the B dimensional attribute descriptor of image $\boldsymbol{I_i}$; the function $\Phi_b : \mathbb{R}^d \mapsto \{0,1\}$ be the b-th attribute detector. Each element in $\mathbf{z_i}$ is determined as:

$$z_{i,b} = \Phi_b(\mathbf{x_i}) \tag{4.1}$$

where $z_{i,b}$ is the b-th element of $\mathbf{z_i}$ and $\mathbf{x_i}$ is the set of features extracted from image $\boldsymbol{I_i}$.

In order to be successful in a classification task, one needs to ensure that the attribute descriptor sufficiently separates images from different categories. Nevertheless, as demonstrated by Farhadi *et al.* in [28], although a set of image attributes can effectively describe objects from different categories, it may not always be sufficient for distinguishing between different categories. This is due to the fact that most image attributes were generated by asking a human to describe images. For instance, it is reasonable to describe a cat as a four legged animal. However, this attribute is not useful to distinguish between cats and dogs as they are both four legged animals. In addition, it is almost impossible to manually identify the subset of discriminative attributes from a large pool of attributes for solving an image classification task with a large number of categories. Therefore, it is important to have an automatic system which is able to identify a subset of discriminative attributes for each application domain.

Another important aspect that should be considered to develop such a system is the fact that the attribute descriptor extraction process is not error-free. This is because the attribute detectors $\{\Phi_b\}_{b=1}^{B}$ are essentially binary classifiers trained to minimise the classification generalisation error. It is preferable to have reliable attribute detectors, which in turn could minimise the overall descriptor extraction error.

In the light of the above facts, we propose that there are intrinsically two aspects contributing to the performance of a classification system utilising image attribute descriptors: (1) attribute discriminative power; and (2) attribute reliability. The former determines the separability between image categories, and the latter determines the reliability of each attribute detector and also the semantic drift of the attribute classifier. Discriminative power has been explored in [123] to discover discriminative category-level attributes. Nevertheless, the discovered attributes resulting from this approach do not necessarily have semantic meaning.

4.2.1 Discriminative and Reliable Attribute Learning

Attribute discriminative power governs how well a set of image attributes separate images from different categories. The attribute discriminative power, Δ can be defined as:

$$\Delta = \sum_{i} \sum_{j} \|\mathbf{z_i} - \mathbf{z_j}\|_H \quad \mathbf{z_i} \in c, \mathbf{z_j} \notin c \tag{4.2}$$

where $\mathbf{z_i}$ and $\mathbf{z_j}$ are the attribute descriptors of the i-th and j-th images which belong to different categories, respectively; $\|\cdot\|_H$ is the hamming distance. The reason why we use hamming distance is because the most popular attribute model is binary model which means the visual attributes are in the form of present or absent situations given a sample. In this chapter we focus on this attribute model. Therefore the visual attribute are in the binary space which relies on hamming distance to describe

the distance between samples. The above equation can be easily extended to the zero-shot learning where only category-level attributes are available:

$$\Delta = \sum_{i} \sum_{j} \|\mathbf{h_i} - \mathbf{h_j}\|_H \quad i \neq j \tag{4.3}$$

where $\mathbf{h_i}, \mathbf{h_j} \in \{0,1\}^B$ are the category-level attribute descriptor. Intuitively, when the attribute discriminative power Δ is maximised, the margin between pair-wise categories will be maximised in the attribute feature space. This will lead to high category separability.

4.2.2 Attribute Reliability

We define the attribute reliability Ω which measures the reliability of a set of attribute detectors as:

$$\Omega = \sum_{b=1}^{B} \omega_b \tag{4.4}$$

where ω_b is the reliability score of the b-th attribute detector. The individual reliability score ω_b is related to the generalisation error of the attribute detector Φ_b. Indeed it is difficult to determine the generalisation error of a classifier [8, 103]. One possible alternative is to define $\hat{\omega_b}$ which is the approximation of ω_b. Thus, the approximated attribute reliability, $\hat{\Omega}$, is defined as:

$$\hat{\Omega} = \sum_{b=1}^{B} \hat{\omega_b}. \tag{4.5}$$

In the present work we determine $\hat{\omega_b}$ by first constructing the Receiver Operating Characteristic (ROC) curve of the attribute detector Φ_b and computing the Area Under the Curve (AUC). We further perform non-linear normalisation using a sigmoid function in order to increase the contrast between the reliable and non-reliable attributes. Therefore, we define $\hat{\omega_b}$ as:

$$\hat{\omega_b} = \frac{1}{1 + e^{-\beta(\mathrm{AUC_b} - \gamma)}} \tag{4.6}$$

where AUC_b is the AUC of the attribute detector Φ_b; β, γ are the normalisation parameters. We determine both the AUC normalisation parameters from a cross-validation set. It is noteworthy to mention the attribute reliability relies on two factors: (1) the generalisation error of the attribute detector; and (2) the semantic drift caused by the noise in the attribute detector training process. The semantic drift happens when an attribute detector accidentally learns a concept different from the initial intention [28]. For instance, when we use car and non-car images as positive and negative samples in order to learn *has wheel* attribute, the corresponding attribute detector may learn *is metallic* concept as the most discriminative feature to differentiate cars with non-car images. Our proposed

approximation of the attribute reliability $\hat{\Omega}$ captures the former factor. Nevertheless, it is still difficult to measure the degree of the attribute detector semantic drift.

4.3 Discriminative and Reliable Attribute Learning

4.3.1 Prior Work

In this part, we will briefly introduce the Discriminative Attribute Prediction (DAP) method proposed in [51]. The DAP uses the Bayes rule to model the relationships between attribute descriptor $\mathbf{z_i}$ and low-level feature representation $\mathbf{x_i}$ of an image as well as $\mathbf{z_i}$ and the unseen test category label v. The attribute descriptor $\mathbf{z_i^y}$ for a seen training category y can be represented as a vector $\left[z_{i,1}^y, ..., z_{i,B}^y\right]^T$, the Bayes posteriori probability for a test category v given an input $\mathbf{x_i}$ can be defined as:

$$P\left(v|\mathbf{x_i}\right) = P\left(v\right)\prod_{b=1}^{B}\frac{P\left(z_{i,b}^v|\mathbf{x_i}\right)}{P\left(z_{i,b}^v\right)} \tag{4.7}$$

where $P(v)$ is the prior of the test category v, $P(z_{i,b}^v)$ denotes the attribute prior, $P(z_{i,b}^v|\mathbf{x_i})$ is the image-attribute probability output of the attribute detector ϕ_b. The authors assume identical test category prior and ignore $P(v)$ effectively. They also use empirical means $P(z_{i,b}^v) = \frac{1}{K}\sum_{k=1}^{K} I(z_{i,b}^k = z_{i,b}^v)$ for all the training categories, where $I(\cdot)$ is the indicator function that gives value one when the condition is met, zero otherwise; and K is the number of training categories. Finally, the best output category from all test categories $v_1, ..., v_q$ is assigned to a test sample $\mathbf{x_i}$ according to the Maximum A Posteriori probability (MAP) prediction as:

$$f\left(\mathbf{x_i}\right) = \underset{q=1,...,Q}{\arg\max}\ P\left(v|x_i\right) = \underset{q=1,...,Q}{\arg\max}\prod_{b=1}^{B}\frac{P\left(z_{i,b}^{v_p}|\mathbf{x_i}\right)}{P\left(z_{i,b}^{v_p}\right)} \tag{4.8}$$

4.3.2 Discriminative and Reliable Attribute Selection

Given a pool of image attributes U, the goal of the present work is to mine the set of attributes which have high discriminative power, as well as reliability. To that end, we define our objective function $\mathrm{J}(\cdot)$ as:

$$\mathrm{J}\left(\{U, \{\mathbf{h_i}\}_{i=1}^{C}, \{\Phi_b\}_{b=1}^{B}\}\right) = \underset{S\in U}{\operatorname{argmax}}\left(\alpha\hat{\Omega}_S + (1-\alpha)\Delta_S\right) \tag{4.9}$$

where $U = \{1 \cdots B\}$ is the set of all image attributes; $S \in U$ is the selected subset of image attributes; Δ_S and $\hat{\Omega}_S$ are the selected attribute discriminative power and attribute reliability, respectively; $\{\Phi_b\}_{b=1}^{B}$ is the set of attribute detectors; C is the number of categories; α is the mixing parameter which determines the importance between attribute discriminative power and reliability.

We note that the optimisation problem presented in (4.9) is NP-hard as it involves optimisation in binary space [71]. This means that the problem cannot be solved by any traditional optimisation algorithm such as gradient descent algorithms. As such, we propose a greedy algorithm wherein for each step, it chooses the attribute that maximises the objective. We call this algorithm Discriminative and Reliable Attribute Learning (DRAL).

The goal of the DRAL algorithm is to select a subset of attributes S so that it maximises $\mathrm{J}(\cdot)$. The algorithm is presented in Algorithm 1. The algorithm optimises the function $\mathrm{J}(\cdot)$ by optimising a single attribute at a time. Let us suppose that we want to optimise the k-th attribute in S. This can be done by converting (4.9) into:

$$\mathrm{J}\left(\{U, \{\mathbf{h_i}\}_{i=1}^{C}, \{\Phi_b\}_{b=1}^{B}\}\right) = \underset{k \in U}{\operatorname{argmax}} \left(\alpha\hat{\omega}_k + (1-\alpha) \sum_i \sum_j \|h_{i,k} - h_{j,k}\|_H + \right.$$
$$\left. \alpha \sum_{b \neq k} \hat{\omega}_b + (1-\alpha) \sum_i \sum_j \sum_{b \neq k} \|h_{i,b} - h_{j,b}\|_H \right) \quad (4.10)$$

which then can be further simplified into:

$$\mathrm{J}\left(\{U, \{\mathbf{h_i}\}_{i=1}^{C}, \{\Phi_b\}_{b=1}^{B}\}\right) = \underset{k \in U}{\operatorname{argmax}} \left(\alpha\hat{\omega}_k + (1-\alpha) \sum_i \sum_j \|h_{i,k} - h_{j,k}\|_H + C \right) \quad (4.11)$$

where $C = \alpha \sum_{b \neq k} \hat{\omega}_b + (1-\alpha) \sum_i \sum_j \sum_{b \neq k} \|h_{i,b} - h_{j,b}\|_H$; $h_{i,k}$ is the k-th element of the category-level attribute descriptor i. To solve the above equation, the proposed algorithm chooses k from U which optimises the above function. Here the k attribute is not included in the set S.

Before optimising the objective function with respect to $k \in U$, we would need to choose $l \in S$ which will be replaced by k. In this case, l needs to be the attribute that is most unreliable and non-discriminative. This means we need to solve the following problem:

$$\mathrm{J}\left(\{U, \{\mathbf{h_i}\}_{i=1}^{C}, \{\Phi_b\}_{b=1}^{B}\}\right) = \underset{l \in S}{\operatorname{argmin}} \left(\alpha\hat{\omega}_l + (1-\alpha) \sum_i \sum_j \|h_{i,l} - h_{j,l}\|_H + C \right) \quad (4.12)$$

where $C = \alpha \sum_{b \neq l} \hat{\omega}_b + (1-\alpha) \sum_i \sum_j \sum_{b \neq l} \|h_{i,b} - h_{j,b}\|_H$. The above equation can be addressed by choosing the attribute from the selected subset S which minimises the function.

Given a subset S, the algorithm will alternate between solving (4.12) and (4.11). It stops when the member of subset S does not change any further.

Algorithm 1: The proposed greedy algorithm for solving (4.9). The final result is S which is the most discriminative and reliable attribute set selected from U; N is the number of attributes (*i.e.,* $N = |S|$)

Input: $\{U, \{\mathbf{h_i}\}_{i=1}^{C}, \{\Phi_b\}_{b=1}^{B}\}, N$

1: $S \leftarrow$ randomly select N number of attributes from U
2: **repeat**
3: $\quad l \in S \leftarrow$ Solve (4.12)
4: $\quad S = S - \{l\}$
5: $\quad k \in U \leftarrow$ Solve (4.11)
6: $\quad S = S \cup \{k\}$
7: **until** S does not change

There are several design choices on how S is initialised. However, from our empirical analysis, initialising S by randomly selecting attributes from U always gives quick convergence. Therefore, we will use random selection to initialise S. The full algorithm is presented in Algorithm 1. We will later show in the experiment that by doing this procedure, the algorithm monotonically increases the objective function and thus convergence can be reached.

As to the computational complexity, first term in (4.9) can be understood as the computational cost of computing AUC, in the other word, it is just the computational complexity of each attribute detector. In this method, we apply binary SVM classifier as the detector. Therefore the computational complexity of first term is $O(max(n,d)min(n,d)^2)$ [14] where n is the number of points and d is the number of dimensions. For the second term, it computes the pairwise hamming distance between samples in the form of a fixed length of attribute representation. Therefore computational complexity of second term is $O(n^2 q(S))$, where $q(S)$ is the number of attributes in set S. For the maximum procedure, we need to another loop to traverse all the attributes in S. Thus the final computational complexity is $O(q(S) * ((max(n,d)min(n,d)^2) + n^2 q(S)))$. The computational complexity is not very high.

Another way to solve (4.9) is by considering a group of attributes instead of an individual attribute. We call this approach as group selection approach. Unlike the proposed approach, in the group selection approach, at one instance, we would like to select a group of attributes that will optimise (4.9). Nevertheless, from our observation, in this setting, the solution can always be reduced to the single attribute selection presented in (4.11) and (4.12). This reduces the group selection approach to give virtually the same results as the proposed approach.

4.4 Experiment Evaluation

In this section, the variants of the proposed approach are evaluated and compared. Then the best performing system will be contrasted to the state-of-the-art method named Direct Attribute Prediction (DAP) [51]. We note that we use the same classifier as DAP for all variants. The difference is that the DAP uses the whole set of attribute pool. We consider the zero-shot learning problem applied in the Animal with Attribute dataset (AwA) [51].

4.4.1 Dataset and Experiment Settings

The AwA dataset contains 35,474 images of 50 animal categories with 85 attribute labels. It has two types of labels for each image: the attribute label and category label. Category label indicates the animal category to which the image belongs. Attribute label represents the presence/absence of an attribute in an image. Therefore, each image is represented by 85-dimensional attribute descriptor. We note that in this dataset, all the images in the same category have the same attribute representation. We follow the experiment protocol and the settings used in [51] for the zero-shot learning problem. In particular, the categories are divided into two disjoint sets: 40 categories for training and 10 categories for testing. In this way, there is no training image given for the 10 categories in the test set. However, manually labelled category-level attributes for each test category are given.

For the low-level feature used to train the attribute detectors and detect the attributes, we use the same extracted features as in [51] such as: HSV colour histogram, SIFT [68], rgSIFT [102], PHOG [7], SURF [3], and local self-similarity [96]. All the features are combined using the Multiple Kernel Learning. We also use the kernels provided in [51], to make our results comparable to the previous works. In addition, we also use the same parameters to train the attribute detectors, and repeat the experiment 5 times.

The proposed DRAL algorithm has three parameters: β and γ which are used for (4.6) and the mixing coefficient α. The values of all parameters are selected from the cross-validation set. From our empirical analysis, we found that $\gamma = \frac{1}{B}\sum_b \hat{\omega}_b$ to be a good value. In addition β is determined from range $[0, 100]$.

The mixing coefficient α determines the importance of the attribute properties (*i.e.,* attribute discriminative power and attribute reliability). We search α with range $[0.1...0.9]$ and we find that $\alpha = 0.9$ to performs the best. Intuitively, we should put more importance on the attribute discriminative power when there are a large number of categories. This can be explained by the fact that a large number of categories require longer binary code to sufficiently separate them. However, as mentioned, there are only 10 categories in the test set; thus, we need to put more importance toward the attribute reliability.

4.4.2 Experimental Results

For the first evaluation, we compare five variants of the proposed system: (1) DRAL using only the attribute discriminative power information (*i.e.,* $\alpha = 0$), denoted DRAL (discriminative); (2) DRAL using only the attribute reliability information (*i.e.,* $\alpha = 1.0$), DRAL (reliability); (3) the proposed DRAL using both attribute properties, denoted DRAL (both); (4) semi-random selection, and (5) random selection. The semi-random selection approach uses the DRAL algorithm without solving (4.12). Instead, the approach randomly selects $l \in S$. Whilst the random selection approach randomly selects S from U.

We first present the empirical study of the study the proposed algorithm's convergence. Figure 4.1 shows the plot of the objective function score presented in (4.9) for each variant of DRAL in every loop. Note that for the case of random selection, the attribute set S is randomly selected for every loop. This result suggests that when using both attribute properties, the proposed algorithm achieves the highest convergence rate (*i.e.,* after iteration 20). Moreover, the other approaches are not able to maximise the objective function. The semi-random selection variant requires more iterations to converge. This shows that our strategy requires both attribute properties in order to maximise the objective function.

It is noteworthy to mention that the algorithm did not converge when using only the attribute discriminative power property. On closer examination, we found that the system picked many unreliable attributes. Although, the system generate discriminative attribute descriptors that sufficiently separate the 10 test categories, the generated large errors (unreliability) during the attribute descriptor extraction on each test image led to a large classification error.

In the second evaluation, we compared the performance of all variants in the test set. To this end, we varied the number of selected attribute, N from 35 to 75. Figure 4.2 presents the results. The proposed DRAL algorithm using both attribute properties generally performs better than the other variants. In Figure 4.1, the variant achieves slightly better performance than the original DAP when only 55 attributes were selected (*i.e.,* 35% less). Moreover, we can reduce this to 40 with a price of slight performance loss (from 41.5% to 41.2%). This suggests that the proposed algorithm is able to select the most discriminative attribute set from the 85 attributes provided in the dataset. Table 4.1 presents further detailed results when the number of attributes was set to 55. These results are consistent with the convergence evaluation presented before.

4.4.3 Comparative Analysis to DAP

In this evaluation, we use the best performing system previously found (*i.e.,* DRAL (both)). Figure 4.3 present the comparison between the DAP and the DRAL ROC curves. This further validates the efficacy of the proposed system. The AUCs of the system in most categories are better than those of the DAP. This suggests that the automatic selection of discriminative and reliable attributes does indeed

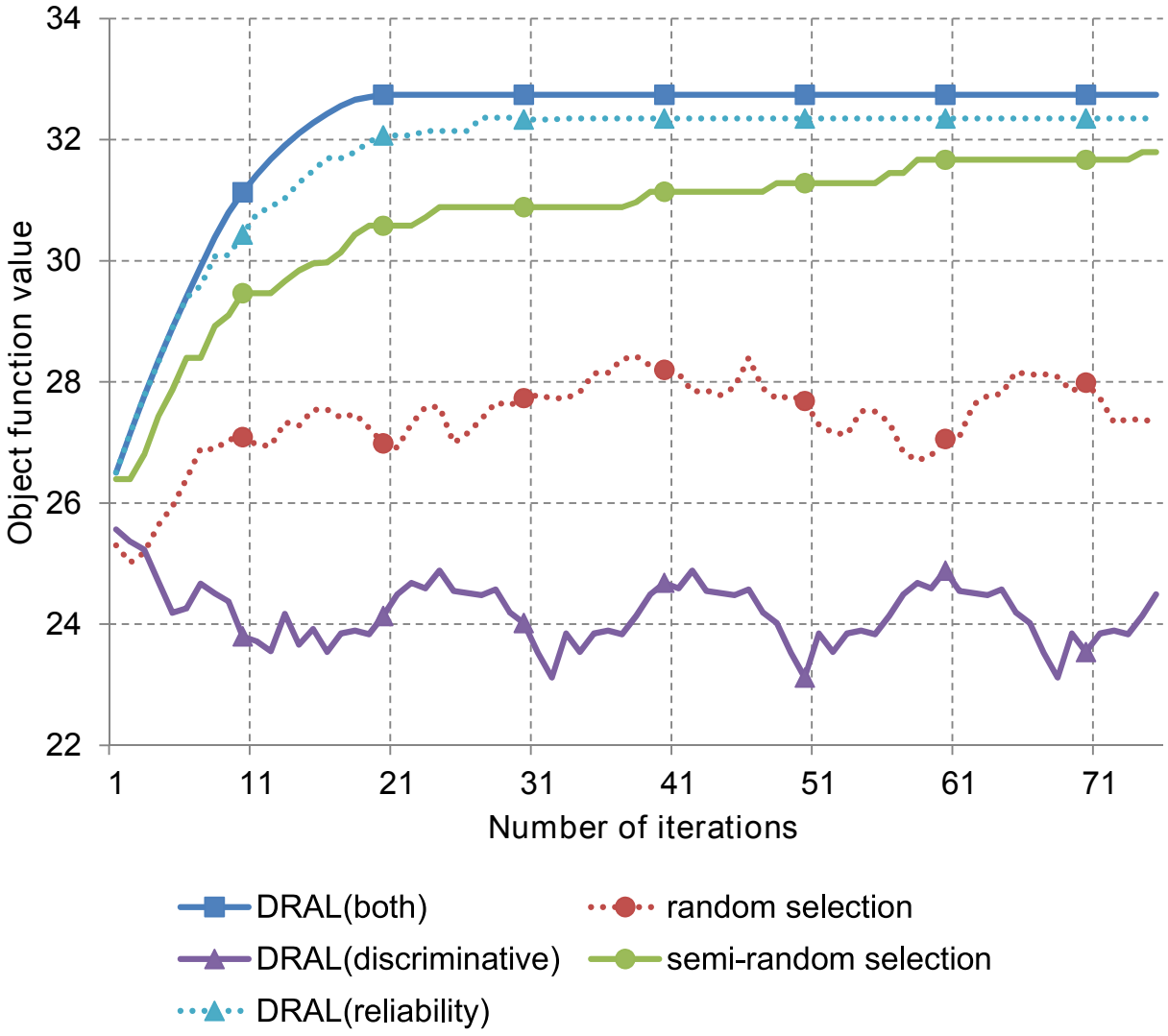


Figure 4.1: The plot of objective function (4.9) for each variant of the proposed approach

Table 4.1: Zero-shot multi-class classification accuracy on 10 novel animal categories selecting 55 attributes.

Methods	**Accuracy (in %)**
DRAL (both)	**41.5**
DRAL (reliability)	40.9
DRAL (discriminative)	37.5
Semi-random selection	40.6
Random selection	40.2
original DAP [51]	41.4

notably improve the performance over the DAP in most test categories. In particular, it significantly

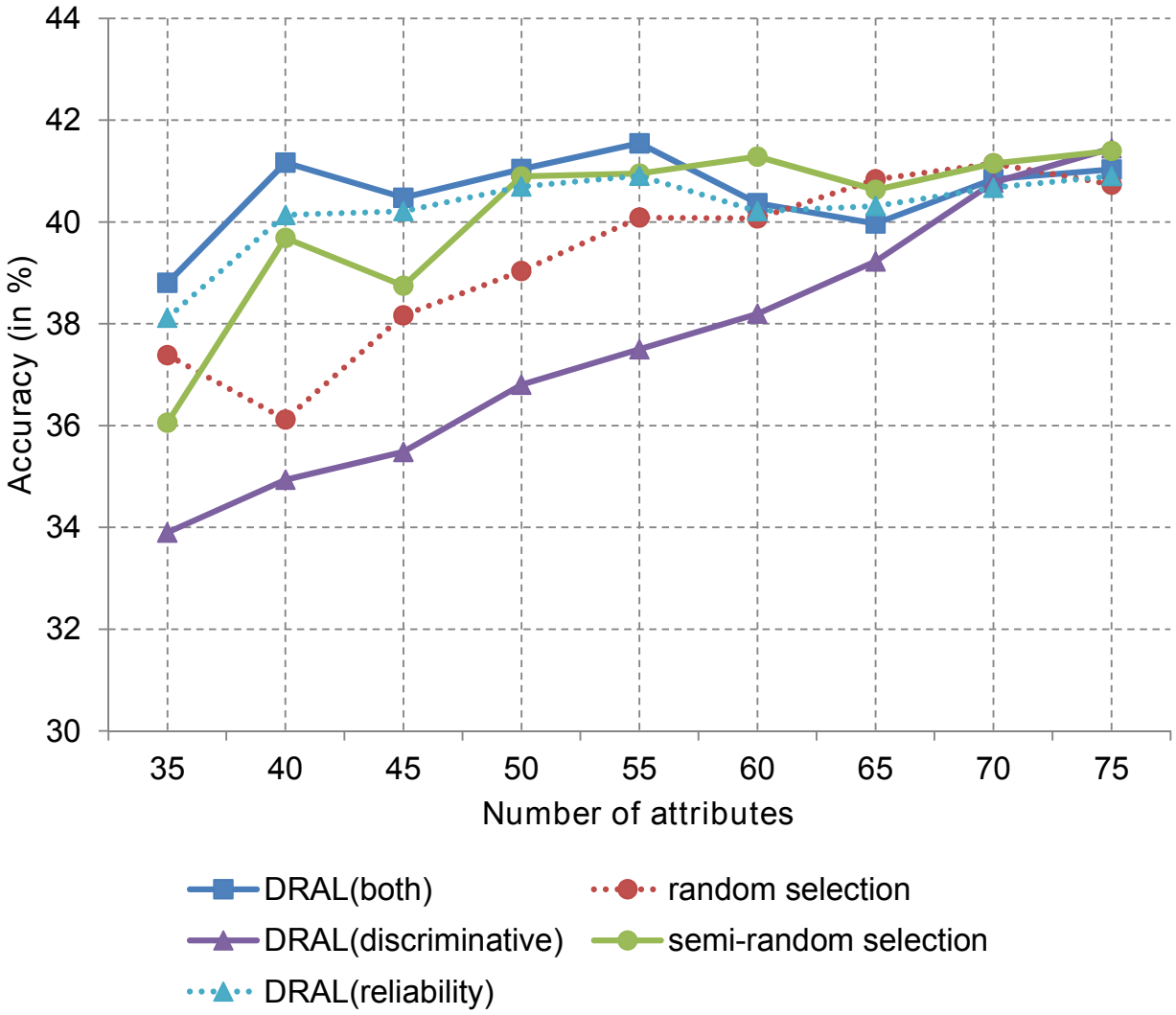


Figure 4.2: Comparison of the proposed approach variants when the number of selected attributes varies from 35 to 75, the best performance (41.5% in accuracy outperforms that of DAP method in Table 4.1) appears at the point when 55 selected attributes used.

outperforms DAP in *leopard*, *persian+cat*, *chimpanzee*, and *seal*. However, we note that there are still two categories performing worse, namely the *pig* and *hippopotamus* categories.

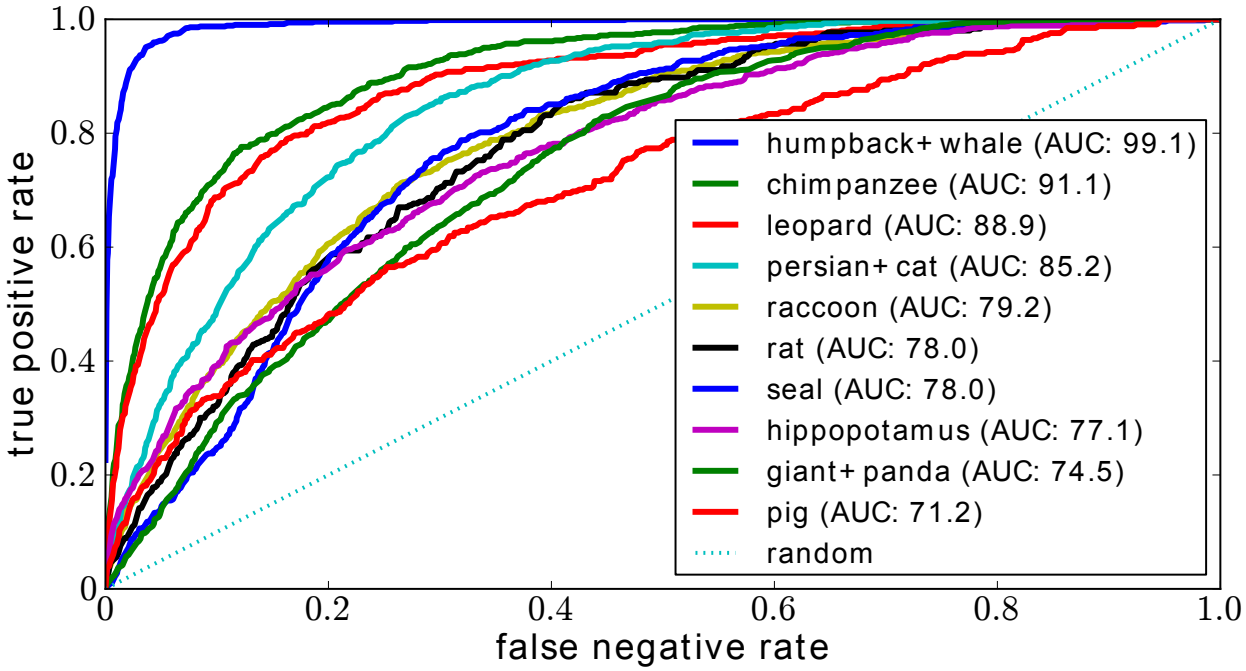
4.5 Summary

Image attributes offer a convenient way of bringing semantic concepts into a machine-readable image representation. Although these image attributes are generally semantically meaningful, they are not necessarily discriminative. This means that there is no guarantee for image classification systems using this approach to achieve good performance. To that end, in the chapter we study two properties of image attributes: attribute discriminative power and attribute reliability. The attribute discrimi-

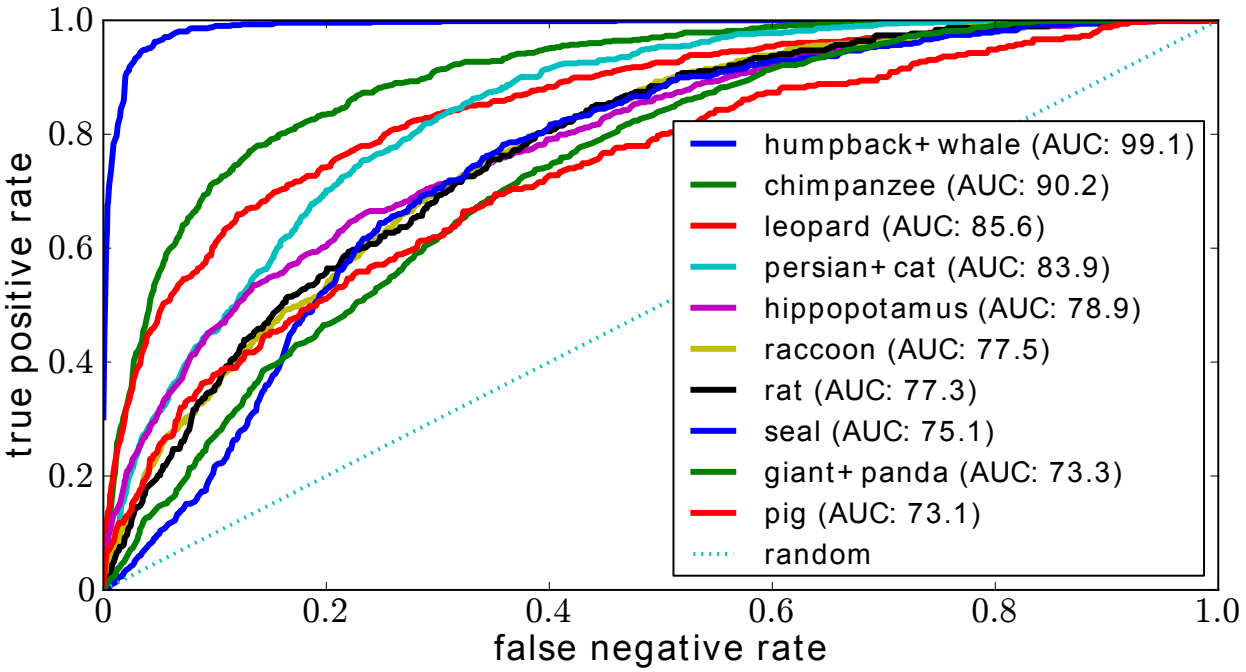
native power is related to the property of a set of image attributes to separate images of different categories. The attribute reliability is related to the error produced during the attribute descriptor extraction process, We propose a greedy algorithm, here denoted Discriminative Reliable Attribute Learning (DRAL), to select a subset of attributes to maximise an objective function that incorporates the two properties. Given a pool of image attributes, the algorithm first selects the image attribute minimising the objective function from the selected set. Then, it replaces the image attribute from the pool with the one maximising the objective function. The process iterates until the selected set does not change.

We empirically showed that the algorithm converges and was able to optimise the objective function. We contrasted our proposed approach with the state-of-the-art approach, denoted DAP, for the zero-shot learning problem in the Animal with Attribute dataset. The results demonstrated that with significantly less number of attributes, the proposed approach achieved a comparable performance to the DAP approach.

There are many extensions and feasible enhancements that can be explored in the future. For instance, we could use a better approximation to measure the attribute reliability property that considers both the detector performance as well as the semantic drift. Another interesting future direction is to find the smallest set of attributes by adding an additional regularisation term in the objective function. We can also explore some novel applications for the proposed strategy such as super resolution [59], 3D reconstruction [127, 128], or anomaly detection in surveillance systems [116]. Here we can use attributes of low-resolution images as the query to collect the high-resolution images which have similar parts to that, then use the patches of the high-resolution images as sources to approximate the patches of low-resolution images and reconstruct the high-resolution images.



(a) DRAL



(b) DAP

Figure 4.3: Comparison of the Performance between the proposed method DRAL and DAP ROC-curves, and AUC value for the ten test classes

Chapter 5

Automatic and Quantitative Evaluation of Attribute Discovery methods

> What makes things memorable is that they are meaningful, significant, colourful.
>
> Joshua Foer

Chapter Summary: *In this chapter, we propose an attribute meaningfulness metric to address the challenging problem of automatic quantitative evaluation of meaningfulness of the attribute sets discovered by various attribute or code generation methods. More importantly, the automatic comparison of attribute meaningfulness between various attribute discovery methods can be performed in this fashion.*

5.1 Introduction

After introducing the discriminative and meaningful attribute selection in the attribute-level supervised case, in this chapter we focus on the discovery of the meaningful and discriminative attributes from another aspect, the attribute-level unsupervised fashion. However, for the purpose of this, we need to discover what attribute meaningfulness is, and how can we evaluate it for the automatically discovered attribute in an unsupervised way. Unfortunately, there is no such automatic way to evaluate the meaningfulness of automatically discovered attributes except the traditional naive human manual evaluation. In this chapter, we target this shortcoming in the literature to propose a novel metric of the meaningfulness of automatically discovered attributes. From this research, we can begin to glean some insights on mechanisms required to extract meaningful attributes/binary codes which serve as the basics for the work in Chapter 7. Through the comprehensive literature review in Chapter 2 and analysis in this chapter, we are able to answer research question 2 in Section 1.1.

Language is one of the most important factors in communication. We would not have been able to write this thesis if there was not any language! Human language has been used for solving computer vision problems such as scene understanding [80], image or video description [58, 88], and image retrieval [65]. The language model helps us to make an effective transfer of domain expert knowledge into machines. Using human language as a tool to explain the object or scene can be quite powerful to address many computer vision problems [28, 49, 29]. For instance, the active learning framework proposed in [46] employs human knowledge to learn better models. Visual attributes are extremely useful as they are: (1) human understandable; (2) machine computable; and (3) shared across classes. For these reasons, recently there are many attribute discovery methods that have been proposed to extract visual/image attributes [4, 86, 45].

One of the biggest challenges in using attribute descriptors is that an enormous amount of training images with attribute labels is required to train attribute classifiers. It is extremely tedious, time-consuming and expensive to label each individual image for every attribute (*e.g.,* if there are 64 attributes, then each image should have at least 64 labels with one label for each attribute). Furthermore, in some specialised domains such as *Ornithology* [114], *Entomology* [109] and cell pathology [115], the human labelling task could be immensely expensive as only a few highly trained experts could conduct such work.

To reduce the workload, automatic attribute discovery methods have been developed [4, 86, 95, 115, 123]. The primary aim of these works is to learn a function that maps the original image feature space into a binary code space wherein each individual bit represents the presence/absence of a visual attribute. These attribute discovery methods are also closely related to hashing methods [34, 54, 113]. The difference is that, unlike automatic attribute discovery approaches, hashing methods focus more on how to significantly reduce the storage demand and computational complexity whilst maintaining system accuracy.

Despite great strides that have been made in this field, there are still some questions left unaddressed: (1) Given the set of attributes/binary codes discovered by a method, are these attributes or binary codes really meaningful? (2) Can we compare these methods by directly observing the discovered attributes? By exploring these questions, we can begin to glean some insights on mechanisms required to extract meaningful attributes/binary codes. We note that the aim of this work is not to propose a new method to discover attributes. Instead, we propose a meaningfulness metric and use this tool to study existing methods.

Gauging "how meaningful" for a given attribute can be an ill-posed problem as there is no obvious *yardstick* for measuring this. Fortunately, it is pointed out by Parikh and Grauman that meaningful attributes may have a *shared structure* [76, 77]. This means, given an attribute feature space, meaningful attributes are likely to be close to each other within a subspace. In Chapter 6, we further study this shared structure and apply our findings to the task of automatic generation of surveillance video descriptions.

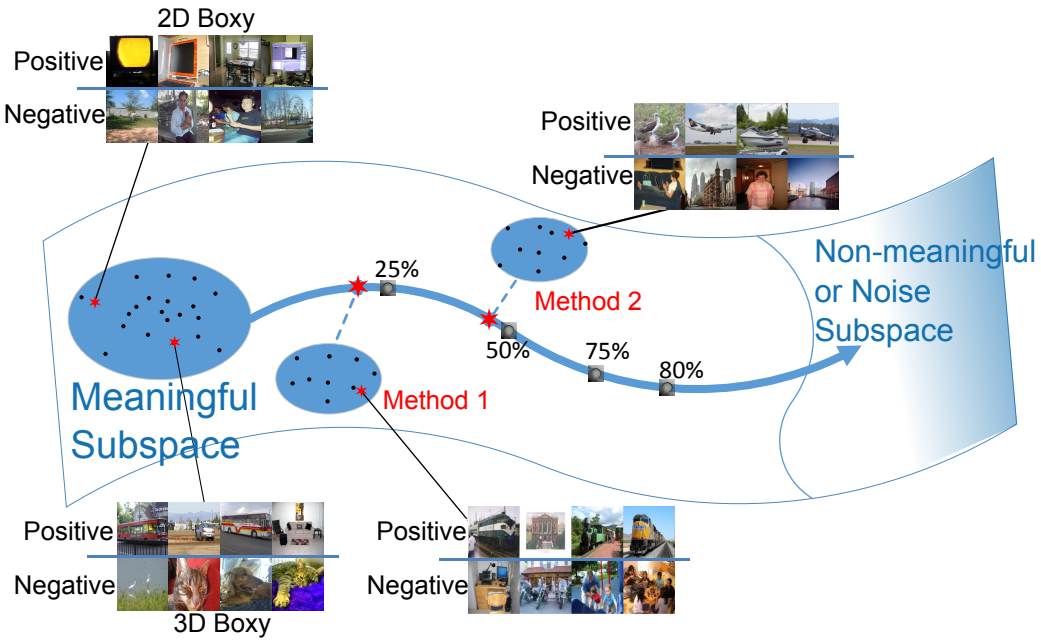


Figure 5.1: An illustration of the proposed attribute meaningfulness metric. Each individual attribute is represented as the outcome of the corresponding attribute classifier tested on a set of images. Inspired by [77], we propose an approach to measure the distance between a set of discovered attributes and the meaningful subspace. The metric score is derived using a subspace interpolation between Meaningful Subspace and Non-Meaningful/Noise Subspace. The score indicates how many meaningful attributes are contained in the set of discovered attributes.

Inspired by previous works, we propose a novel meaningfulness metric that will become one of the *yardsticks* to measure attribute meaningfulness. More specifically, we first measure the distance between the discovered attributes and the meaningful attribute subspace. To this end, an approximate geodesic distance based on reconstruction error is proposed. As it may be difficult to perform quantitative analysis/study using this distance directly, we then derive the meaningfulness metric based on the distance. In particular, the metric is derived by first performing a subspace interpolation between meaningful subspace and non-meaningful subspace, lying on the manifold of decision boundaries. The distance on each interpolated subspace is calculated. These are then used to calibrate the distance of the discovered attributes to the meaningful subspace. Figure 5.1 illustrates our main idea.

Contributions — We list our contributions as follows:

- We propose a reconstruction error based approach with two different regularisations (*i.e.,* ℓ_0 and convex hull) to approximate the geodesic distance between a given attribute set and the Meaningful Subspace.
- We propose the *attribute meaningfulness* metric that allows us to quantitatively measure the meaningfulness of a set of automatically discovered attributes. The metric score is related to "the percentage of meaningful attributes contained in the set of attributes."

- We propose an improved calibration method to avoid pathological cases where the calibration could not be performed. This method is developed based on the in-depth analysis performed in this chapter.
- We present extensive experiments and analysis on four popular attribute datasets to demonstrate that our proposal can really capture attribute meaningfulness. The *attribute meaningfulness* of some recent automatic attribute discovery methods and various hashing approaches are also evaluated on these datasets. A user study is conducted to further validate the effectiveness of the proposed metric. In addition, we compare the proposed metric with a metric adapted from a recent semi-supervised attribute discovery method using the Mixture of Probabilistic Principal Component Analysis (MPPCA) [77, 99]

In this chapter, we perform in-depth analysis on the proposed metric and consider well-known public datasets. Throughout these experiments and in-depth analysis, we found that the calibration step heavily depends on the meaningful subspace spanned by the selected meaningful attributes, denoted as the subspace bases. The space spanned by these bases should be maximised in order to ensure the calibration is performed correctly? To remedy this, we present a simple-yet-effective technique using semantic reasoning and threshold setting.

We continue the chapter as follows. We introduce our approach of evaluating attribute meaningfulness in Section 5.2. Our proposed metric is described in Section 5.3. Next, we discuss the experiments and results in Section 5.4. Finally, the main findings and future directions are presented in Section 5.5.

5.2 Measuring Attribute Set Meaningfulness

We begin by describing the manifold of decision boundaries and the meaningful attribute subspace wherein our proposed metric is based. Then, we define the distance between the automatically discovered attributes and the meaningful attribute set in the manifold space to measure the attribute meaningfulness.

5.2.1 Manifold of Decision Boundaries

Suppose there is a set of samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, an attribute can be considered as a decision boundary which partitions the set into two subsets $\mathcal{X}^+ \cup \mathcal{X}^- = \mathcal{X}$. Here $\mathcal{X}^+$ represents the set where the attribute is present and $\mathcal{X}^-$ represents the set where the attribute is absent. Therefore, all the attributes are lying on a manifold formed by decision boundaries [77].

In this case, an attribute can also be viewed as an N-dimensional binary vector whose element represents the classification output of sample $\mathbf{x}_i$ classified by the corresponding attribute binary classifier $\phi(\cdot) \in \mathbb{R}$. The sign of the classifier output on $\mathbf{x}_i$ indicates whether the sample belongs to the positive

or negative set (*i.e.,* $\mathcal{X}^+$ or $\mathcal{X}^-$). As such, an attribute can be represented as $\mathbf{z}^{[\mathcal{X}]} \in \{-1, +1\}^N$ whose i-th element is $\mathbf{z}^{[\mathcal{X}]}_{(i)} = \text{sign}(\phi(\mathbf{x}_i)) \in \{-1, +1\}$. For the sake of simplicity, we drop the symbol $[\mathcal{X}]$ from $\mathbf{z}^{[\mathcal{X}]}$ whenever the context is clear. Thus, the manifold of decision boundaries w.r.t. $\mathcal{X}$ can be defined as $\mathcal{M}^{[\mathcal{X}]} \in \{-1, +1\}^N$ which is embedded in an N-dimensional binary space. Again, we also write $\mathcal{M}^{[\mathcal{X}]}$ as $\mathcal{M}$.

As observed from [77, 76], the meaningful attributes have shared structure wherein they lie close to each other on the manifold. In other words, all the meaningful attributes are contained in a subspace on $\mathcal{M}$. In an ideal case, all possible meaningful attributes should be included in the subspace. Unfortunately, it is infeasible to enumerate all of them. One intuitive solution is to rely on the existing human knowledge, that is, the human labelled attributes from various datasets such as [6, 76, 77]. These attributes are all naturally meaningful since they are collected through human annotation from the AMT. However, the number of available labelled attributes may not be sufficient. To this end, based on the shared structure assumption, we thus introduce an approximation of the meaningful subspace by linear combinations of the human labelled attributes. This means, if an automatically discovered attribute is close enough to any attribute existing in the meaningful subspace, it should be considered as a meaningful attribute.

5.2.2 Distance of an Attribute to the Meaningful Subspace

In this section, we mathematically define the reconstruction error based distance of an attribute to the Meaningful Subspace. Given a set of N images $\mathcal{X}$, we denote $\mathcal{S} = \{\mathbf{h}_j\}_{j=1}^J, \mathbf{h}_j \in \{-1, +1\}^N$ as the set of meaningful attributes. We use a matrix $\boldsymbol{A} \in \mathbb{R}^{N \times J}$, in which each column vector is the representation of an attribute, to form the set $\mathcal{S}$. As the assumption in [77], meaningful attributes should close to the meaningful subspace spanned by the set of meaningful attributes $\mathcal{S}$. For instance, the primary colours red, green and blue, are able to construct the set of secondary colours such as yellow, magenta, and cyan. Moreover, the primary colours can provide negative information clues to describe other primary colours (*e.g.,* blue is neither green nor red). Under this assumption, we are able to define a reconstruction error based distance between an attribute and the meaningful subspace. More specifically, let $\mathbf{z}_k$ be an attribute and $\boldsymbol{A}$ be the representation of meaningful attributes. The distance is defined as:

$$\min_{\mathbf{r}} \|\boldsymbol{A}\mathbf{r} - \mathbf{z}_k\|_2^2, \tag{5.1}$$

where $\mathbf{r} \in \mathbb{R}^{J \times 1}$ is the reconstruction coefficient vector. Note that the reconstruction in (5.1) may not lie in the manifold $\mathcal{M}$ (i.e. $\boldsymbol{A}\mathbf{r} \notin \mathcal{M}$). Therefore, we relax this reconstruction procedure into Euclidean space for computational simplicity. This relaxation effectively becomes an approximation of the true geodesic distance.

5.2.3 Distance Between a Set of Discovered Attributes and the Meaningful Subspace

Analogously, suppose there are K discovered attributes, we use matrix $\boldsymbol{B} \in \{0,1\}^{N\times K}$ to represent the discovered attribute set $\mathcal{D}$. Then, according to the specific set of images $\mathcal{X}$, we can define the distance between the set of discovered attributes $\mathcal{D}$ and the Meaningful Subspace $\mathcal{S}$ as the average reconstruction error:

$$\delta(\mathcal{D},\mathcal{S};\mathcal{X}) = \frac{1}{K}\min_{\boldsymbol{R}} \|\boldsymbol{A}\boldsymbol{R}-\boldsymbol{B}\|_F^2, \tag{5.2}$$

where $\|\cdot\|_F$ and $\boldsymbol{R} \in \mathbb{R}^{J\times K}$ are the Frobenious norm and the reconstruction matrix respectively.

The reconstruction coefficients are preferably sparse, because generally only a few attributes can provide useful clues to reconstruct a particular attribute, while most of them should stay inactive in this procedure. Similar to the example in Section 5.2.2, only a few colour attributes can reconstruct another colour attribute, most of them should stay inactive (*i.e.,* their reconstruction coefficient should be 0). Unfortunately, the distances in (5.1) and (5.2) may create dense reconstruction coefficients due to the absence of a regularisation term. As such, we first introduce convex hull regularisation. Moreover, according to [94], the perception mechanism of human visual systems follows the sparsity principle. That means only a few attributes will first trigger the semantic-visual connection in our brain. Desirable attribute discovery methods should also obey this principle. Hereby, we consider the sparsity-inducing ℓ_0 regularisation as the second regularisation alternative.

Convex hull regularisation

Via introducing a convex hull constraint, (5.2) becomes:

$$\begin{aligned}\delta_{\text{cvx}}(\mathcal{D},\mathcal{S};\mathcal{X}) = \frac{1}{K}\min_{\boldsymbol{R}} \|\boldsymbol{A}\boldsymbol{R}-\boldsymbol{B}\|_F^2 \,\text{s.t.}\\ \boldsymbol{R}(i,j) \geq 0\\ \sum_{i=1}^{J}\boldsymbol{R}(i,\cdot) = 1.\end{aligned} \tag{5.3}$$

This objective function describes the average distance between each discovered attribute $\mathbf{z}_k \in \mathcal{D}$ and the convex hull of $\mathcal{S}$. Its optimisation can be efficiently solved using the method proposed in [10].

ℓ_0 regularisation

As to ℓ_0 regularisation, different from the convex hull regularisation, a possible direct correlation between each discovered attribute $\mathbf{z}_k \in \mathcal{D}$ and the meaningful attribute, $\mathbf{h}_j \in \mathcal{S}$ is considered:

$$\begin{aligned}\delta_{\mathrm{jp}}(\mathcal{D},\mathcal{S};\mathcal{X}) =& \frac{1}{K}\min_{\boldsymbol{R}}\|\boldsymbol{A}\boldsymbol{R}-\boldsymbol{B}\|_F^2, \text{s. t.} \\ & \forall k\in\{1\cdots K\}, \|\boldsymbol{R}_{\cdot,k}\|_0\leq 1, \\ & \forall j\in\{1\cdots J\}, \|\boldsymbol{R}_{j,\cdot}\|_0\leq 1.\end{aligned} \tag{5.4}$$

where $\boldsymbol{R}_{j,\cdot}$, $\boldsymbol{R}_{\cdot,k}$ represent the j-th row vector and the k-th column vector in matrix $\boldsymbol{R}$ respectively. The two additional ℓ_0 regularisers enforce one-to-one relationships between $\mathcal{S}$ and $\mathcal{D}$. The reconstruction matrix $\boldsymbol{R}$ correlates each individually discovered attribute to each meaningful one. Namely, for each discovered attribute $\mathbf{z}_k \in \mathcal{D}$, the closest $\mathbf{h}_j \in \mathcal{S}$ is found to minimise the function. However, it could be possible that $|\mathcal{S}| > |\mathcal{D}|$. In this case, we can only match K discovered attributes in $\mathcal{S}$ and vice-versa.

Unfortunately, the optimisation for (5.4) is non-convex. As such, a greedy approach is proposed to address this through iteratively finding pairs of meaningful discovered attributes with the smallest distance. This can be converted into finding the pairs with the highest similarities (lowest distance means highest similarity).

Here we can define the similarities between a discovered attribute $\mathbf{z}_k$ and a meaningful attribute $\mathbf{h}_j$ in terms of their correlations. Let $\rho(\mathbf{z}_k, \mathbf{h}_j), \mathbf{z}_k \in \mathcal{D}, \mathbf{h}_j \in \mathcal{S}$ be the correlation between $\mathbf{z}_k$ and $\mathbf{h}_j$. Then ρ can be defined as:

$$\rho(\mathbf{z}_k, \mathbf{h}_j) = \frac{\mathrm{count}(\mathbf{z}_k = \mathbf{h}_j)}{N}, \tag{5.5}$$

where count means the operation which counts how many same elements $\mathbf{z}_k$ shares with $\mathbf{h}_j$.

Thus, the function $\rho(\mathbf{z}_k, \mathbf{h}_j)$ can be computed from $\boldsymbol{A}_{\cdot,j}$ and $\boldsymbol{B}_{\cdot,k}$, where $\boldsymbol{A}_{\cdot,j}$, $\boldsymbol{B}_{\cdot,k}$ represent the discovered attribute $\mathbf{z}_k$ and the meaningful attribute $\mathbf{h}_j$ respectively. Denote $\mathcal{P}$ as the set of M pairs of $\mathbf{h}_j \in \mathcal{S}$ and $\mathbf{z}_k \in \mathcal{D}$ that have the highest correlation, $\mathcal{P} = \{(\mathbf{h}_j^1, \mathbf{z}_k^1)\cdots(\mathbf{h}_j^M, \mathbf{z}_k^M)\}$, $\mathbf{h}_j^i = \mathbf{h}_j^l$ if and only if $i = l$, $\mathbf{z}_k^i = \mathbf{z}_k^l$.

Therefore the matrix $\boldsymbol{R}^*$ that minimises (5.4) is defined, after $\mathcal{P}$ is determined, via:

$$\boldsymbol{R}^*_{j,k} = \begin{cases} 1 \text{ if } (\mathbf{h}_j, \mathbf{z}_k) \in \mathcal{P} \\ 0 \text{ if } (\mathbf{h}_j, \mathbf{z}_k) \notin \mathcal{P}. \end{cases} \tag{5.6}$$

For the given inputs $\mathcal{D} = \{\mathbf{z}_k\}_{k=1}^K$, $\mathcal{S} = \{\mathbf{h}_j\}_{j=1}^J$ and $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, Algorithm 2 elaborates the procedures of computing the set $\mathcal{P}$. Note that, $(\mathbf{h}_j, \cdot)$ and $(\cdot, \mathbf{z}_k)$ in step 3 represent all possible pairs containing $\mathbf{h}_j$ and $\mathbf{z}_k$, respectively.

Algorithm 2: The proposed greedy algorithm to solve (5.4)

Input: $\mathcal{D} = \{\mathbf{z}_k\}_{k=1}^{K}$, $\mathcal{S} = \{\mathbf{h}_j\}_{s=1}^{J}$ and $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$
Ensure: $\mathcal{P}$ that contains M pairs that have the highest correlation, where $M = \min(K, J)$.
1: $\mathcal{P} \leftarrow \{\}$
2: **while** $|\mathcal{P}| \leq M$ **do**
3: Find the highest $\rho(\mathbf{h}_j, \mathbf{z}_k)$ where $(\mathbf{h}_j, \cdot) \notin \mathcal{P}$ and $(\cdot, \mathbf{z}_k) \notin \mathcal{P}$.
4: $\mathcal{P} = \mathcal{P} \cup (\mathbf{h}_j, \mathbf{z}_k)$
5: **end while**

5.3 Attribute Set Meaningfulness Metric

Attribute meaningfulness metric is designed to determine which existing automatic attribute discovery method is more likely to discover meaningful attributes. Moreover, it can be provide some insights about how to devise new automatic attribute discovery methods.

In this section we will introduce the Attribute Set Meaningfulness Metric. We order our discussion as follows: (1) Meaningful subspace interpolation, (2) Selecting meaningful subspace representation, and (3) Computing the meaningfulness metric.

By means of the distance functions δ_{jp} and δ_{cvx} described in Section 5.2.2, we are able to measure how far the set of discovered attributes $\mathcal{D}$ is from the Meaningful Subspace $\mathcal{S}$. The closer the distance, the more meaningful the set of attributes are. However, as the relationship between the proposed distances and meaningfulness could be non-linear, the distance may not be easy to interpret. Furthermore, it is difficult to compare the results between δ_{cvx} and δ_{jp}.

5.3.1 Attribute Meaningful Subspace Interpolation

Our goal is to obtain a metric that is both easy to interpret and able to perform comparisons between various distance functions. Inspired by [35], we apply the subspace interpolation to generate a set of subspaces between Meaningful Subspace, and Non-Meaningful Subspace or Noise Subspace. Here, we use a set of evenly distributed random attributes to represent the Non-Meaningful Subspace $\mathcal{N}$.

For the purpose of subspace interpolation, the meaningful attribute set $\mathcal{S}$ is divided into two subsets:

$$\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S} \tag{5.7}$$

where, we consider the set $\mathcal{S}^1$ as the representation of the Meaningful Subspace. When gradually adding random attributes $\tilde{\mathcal{N}} \in \mathcal{N}$ into $\mathcal{S}^2$, the interpolated sets of subspaces can be obtained. Here we present the proposition, which guarantees that the interpolation is able to generate subspaces between the Meaningful Subspace and the Non-Meaningful Subspace.

Proposition 5.3.1. *Let $\tilde{\mathcal{S}} = \mathcal{S}^2 \cup \tilde{\mathcal{N}}$; when $\tilde{\mathcal{N}} = \{\}$, the distance δ^* between $\tilde{\mathcal{S}}$ and $\mathcal{S}^1$ (refer to* (5.7)*) is minimised. However, when $\tilde{\mathcal{N}} \to \mathcal{N}$, the distance between $\tilde{\mathcal{S}}$ and $\mathcal{S}^1$ is asymptotically close to $\delta^*(\mathcal{S}^1, \mathcal{N}; \mathcal{X})$, where δ^* is the distance function presented previously such as δ_{jp} and δ_{cvx}. More precisely, we can define the relationship as follows:*

$$\lim_{|\tilde{\mathcal{N}}| \to \infty} \delta^*(\tilde{\mathcal{S}}, \mathcal{S}^1; \mathcal{X}) = \delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X}). \tag{5.8}$$

Proof. Let $\boldsymbol{R}^*$ be the solution for the distance δ^*. The distance δ^* can be computed as follows:

$$\begin{aligned}
\delta^*(\mathcal{S}^2 \cup \tilde{\mathcal{N}}, \mathcal{S}^1; \mathcal{X}) &= \frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \|\boldsymbol{A}\boldsymbol{R}^* - \boldsymbol{B}\|_F^2 \\
&= \frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \sum_{\mathbf{b}_i \in \mathcal{S}^2 \cup \tilde{\mathcal{N}}} \|\boldsymbol{A}\mathbf{r}_i^* - \mathbf{b}_i\|_2^2 \\
&= \frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \left\{ \sum_{\mathbf{b}_j \in \mathcal{S}^2} \|\boldsymbol{A}\mathbf{r}_j^* - \mathbf{b}_j\|_2^2 + \sum_{\mathbf{b}_l \in \tilde{\mathcal{N}}} \|\boldsymbol{A}\mathbf{r}_l^* - \mathbf{b}_l\|_2^2 \right\}
\end{aligned} \tag{5.9}$$

As $\mathcal{S}^2$ is assumed to be meaningful and $\tilde{\mathcal{N}}$ is not, then adding attribute $\mathbf{b}_l$, a member of $\tilde{\mathcal{N}}$, should increase the average distance. Thus, we have the following:

$$\frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \left\{ \sum_{\mathbf{b}_j \in \mathcal{S}^2} \|\boldsymbol{A}\mathbf{r}_j^* - \mathbf{b}_j\|_2^2 + \sum_{\mathbf{b}_l \in \tilde{\mathcal{N}}} \|\boldsymbol{A}\mathbf{r}_l^* - \mathbf{b}_l\|_2^2 \right\} \geq \frac{1}{|\mathcal{S}^2|} \sum_{\mathbf{b}_j \in \mathcal{S}^2} \|\boldsymbol{A}\mathbf{r}_j^* - \mathbf{b}_j\|_2^2 \tag{5.10}$$

It means that the distance between $\tilde{\mathcal{S}} = \mathcal{S}^2 \cup \tilde{\mathcal{N}}$ and $\mathcal{S}^1$ can only be minimised when $\tilde{\mathcal{N}}$ is an empty set. On the other hand, when we keep increasing the size of $\tilde{\mathcal{N}}$, the contribution of the second term in (5.9) becomes more significant than the first term. Thus, (5.9) is approximately close to:

$$\approx \frac{1}{|\tilde{\mathcal{N}}|} \left\{ \sum_{\mathbf{b}_l \in \tilde{\mathcal{N}}} \|\boldsymbol{A}\mathbf{r}_l^* - \mathbf{b}_l\|_2^2 \right\}, \text{ as } |\tilde{\mathcal{N}}| >> |\mathcal{S}^2| \tag{5.11}$$

In addition, as $|\tilde{\mathcal{N}}| \to \infty$, $\tilde{\mathcal{N}}$ will be close to $\mathcal{N}$. Thus, the above equation is approximately close to:

$$\begin{aligned}
&\approx \frac{1}{|\mathcal{N}|} \left\{ \sum_{\mathbf{b}_l \in \mathcal{N}} \|\boldsymbol{A}\mathbf{r}_l^* - \mathbf{b}_l\|_2^2 \right\} \\
&= \delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X})
\end{aligned} \tag{5.12}$$

□

Remarks. Proposition 5.3.1 basically describes when random attributes are added into $\tilde{\mathcal{S}}$ gradually, the subspace that is initially close to the Meaningful Subspace $\mathcal{S}^1$ will be increasingly more distant

from $\mathcal{S}^1$. Eventually the subspace will be spanned by random attributes that is asymptotically close to the Non-Meaningful attribute subspace.

5.3.2 Selecting Meaningful Subspace Representation

As discussed in Section 5.2.1, enumerating all of the meaningful attributes to represent the meaningful attribute subspace is impossible. We thus use linear combinations of meaningful attributes to approximate the meaningful subspace.

However, the division of the meaningful attributes into two subsets, as suggested in Proposition 5.3.1, will reduce the subspace spanned to represent the meaningful subspace. More specifically, the linear combination of attributes from $\mathcal{S}^1$ may not span the whole meaningful subspace. To remedy this, one should carefully select the attributes to form $\mathcal{S}^1$ that can maximise the space spanned by the representation.

Under our proposed approach, which is based on the linear reconstruction, the selected meaningful attributes for $\mathcal{S}^1$ should form the basis of the meaningful subspace. Here, one way to maximise the space spanned is to select independent bases.

The attribute independence with respect to the others can be evaluated by how well the attribute can be reconstructed from others. In addition, one can evaluate the attribute independence from the attribute semantic names. For instance, textural attributes such as 'metal' may be independent to the other textural attributes such as 'grass' or 'wooden'. Therefore, these attributes should be included in the set to represent the meaningful subspace *i.e.,* the set $\mathcal{S}^1$.

In light of these facts, we propose an approach to perform meaningful subspace representation selection, $\mathcal{S}^1$. First, the attribute semantic descriptions are considered. Any attributes that are deemed independent will be indicated and always put in the set $\mathcal{S}^1$. On the second step, we evaluate the attribute independence by applying either δ_{cvx} or δ_{jp}. In particular, we use a leave-one-attribute-out scheme, which calculates the distance between one attribute to the rest of the attributes. We then set the threshold α. Again, we indicate any attributes having distance more than α and always put them in $\mathcal{S}^1$. The threshold, α will be one of the parameters, which will be determined during the experiments.

5.3.3 Computing the Meaningfulness Metric

After constructing the meaningful subspace, we can calibrate the attribute set meaningfulness distance by subspace interpolation based on the equivalent distance effect assumption [5]. That is, if the distance of two attribute subspaces to the meaningful subspace are the same, the amount of meaningful contents contained in these two subspaces is the same.

We denote the distance between $\tilde{\mathcal{S}}$ and the Meaningful Subspace $\mathcal{S}^1$ as $\delta^{\tilde{\mathcal{S}}}$ and the distance between $\mathcal{D}$ and the Meaningful Subspace $\mathcal{S}^1$ as $\delta^{\mathcal{D}}$. After subspace interpolation, we find the subspace $\tilde{\mathcal{S}}$ that

makes $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$. Using the equivalent distance effect assumption, if $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$, the meaningfulness between $\tilde{\mathcal{S}}$ and $\mathcal{D}$ should be on par with each other. As $\tilde{\mathcal{S}}$ is defined as a set of meaningful attributes added with additional noise attributes, this representation is able to evaluate the meaningfulness of $\mathcal{D}$. We can consider this task as an optimisation problem as follows:

$$g^* = \arg\min_{|\tilde{\mathcal{N}}|} \left\| \delta^*(\{\mathcal{S}^2 \cup \tilde{\mathcal{N}}\}, \mathcal{S}^1; \mathcal{X}) - \delta^*(\mathcal{D}, \mathcal{S}^1; \mathcal{X}) \right\|_2^2 \tag{5.13}$$

where g^* represents the minimum number of random attributes required to be added into $\tilde{\mathcal{S}}$ to make $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$. The above optimisation problem can be interpreted as searching for the furthest subspace $\tilde{\mathcal{S}}$ from the Meaningful Subspace in an open sphere with radius $\delta^{\mathcal{D}}$. The above equation can be simply solved by a curve fitting approach. In our implementation, we apply the least square approach.

Finally, we denote γ as the proposed attribute meaningfulness metric as follows:

$$\gamma(\mathcal{D}; \mathcal{X}, \mathcal{S}) = (1 - \frac{g^*}{|\mathcal{S}^2| + g^*}) \times 100. \tag{5.14}$$

Remarks. The equation in (5.14) determines how many noise/Non-Meaningful attributes are required for a set of automatically discovered attributes to have similar distance as $\delta^{\mathcal{D}}$. On the other hand, our proposed metric reflects how many meaningful attributes are contained in the attribute set. A smaller number of Non-Meaningful attributes indicates a more meaningful attribute set overall.

Since different aspects of meaningfulness may be captured by various distance functions, we combine the metric values calculated using both proposed distance functions. For simplicity, we use an equally weighted summation: $\tilde{\gamma} = \frac{1}{2}\gamma_{\text{cvx}} + \frac{1}{2}\gamma_{\text{jp}}$, as our final metric where γ_{cvx} and γ_{jp} are metric score results when (5.14) are respectively applied to reconstruction error with convex hull regularisation and with ℓ_0 regularisation.

5.4 Experiments

In this part, the efficacy of our approach to measure the meaningfulness of a set of attributes will be first evaluated. Then the proposed metric is used to evaluate meaningfulness of the attribute sets generated by various automatic attribute discovery methods such as PiCoDeS [4] and Discriminative Binary Codes (DBC) [86] as well as some recent hashing methods such as Iterative Quantisation (ITQ) [34], Spectral Hashing (SPH) [113], Locality Sensitivity Hashing (LSH) [54] and Kernel-Based Supervised Hashing (KSH) [64].

The two proposed metrics γ_{jp} (5.4), γ_{cvx} (5.3) and the combined metric $\tilde{\gamma}$ are applied to evaluate the meaningfulness of the attributes discovered from the comparative methods on four attribute datasets: (1) a-Pascal a-Yahoo dataset (ApAy) [28]; (2) Animal with Attributes dataset (AwA) [51];

(3) SUN Attribute dataset (ASUN) [80]; (4) Unstructured Social Activity Attribute dataset (USAA) [33]

Finally, our metric will then be compared against a user study and a metric, denoted the MPPCA metric or MPPCA, adapted from the semi-supervised attribute discovery method proposed in [77].

5.4.1 Datasets and Experiment Setup

a-Pascal a-Yahoo dataset (ApAy) [28] — comprises two sources: a-Pascal and a-Yahoo. There are 12,695 cropped images in a-Pascal that are divided into 6,340 for training and 6,355 for testing with 20 categories. The a-Yahoo set has 12 categories disjoint from the a-Pascal categories. Moreover, it only has 2,644 test exemplars. There are 64 attributes provided for each cropped image. The dataset provides four features for each exemplar: local texture, HOG, edge and colour descriptor. We use the training set for discovering attributes and we perform our study on the test set. More precisely, we consider the test set as the set of images $\mathcal{X}$ defined in Section 5.2.1.

Animal with Attributes dataset (AwA) [51] — the dataset contains 35,474 images of 50 animal categories with 85 attribute labels. There are six features provided in this dataset: HSV colour histogram, SIFT [68], rgSIFT [102], PHOG [7], SURF [3], and local self-similarity [96]. The AwA dataset is proposed for studying the zero-shot learning problem. As such, the training and test categories are disjoint; there are no training images for test categories and vice-versa. More specifically, the dataset contains 40 training categories and 10 test categories. Similar to the ApAy dataset, we use the training set for discovering attributes and we perform the study in the test set.

SUN Attribute dataset (ASUN) [80] — ASUN is a fine-grained scene classification dataset consisting of 717 categories (20 images per category) and 14,340 images in total with 102 attributes. There are four types of features provided in this dataset: (1) GIST; (2) HOG; (3) self-similarity; and (4) geometric context colour histograms (See [119] for feature and kernel details). From 717 categories, we randomly select 144 categories for discovering attributes. As for our evaluation, we randomly select 1,434 images (*i.e.,* 10% of 14,340 images) from the dataset. It means, in our evaluation, some images may or may not come from the 144 categories used for discovering attributes.

Unstructured Social Activity Attribute dataset (USAA) [33] — USAA is a relatively novel benchmark attribute dataset for social activity video classification and annotation. It is manually annotated with 69 groundtruth attributes from 8 semantic class videos from Columbia Customer Video (CCV) dataset. There are 100 videos per-class for training and testing respectively. The annotated attributes can be divided into 5 broad categories: actions, objects, scenes, sounds, and camera movement. The 8 classes in the dataset are birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance, and wedding reception. The SIFT, Space-Time Interest Points (STIP), and Mel-scale Frequency Cepstral Coefficient (MFCC) features for all these videos are extracted in the dataset.

For each experiment, we apply the following pre-processing step described in [4]. We first lift each feature into a higher-dimensional space, which is three times larger than the original space. After the features are lifted, we then apply Principle Component Analysis (PCA) to reduce the dimensionality of the feature space by 40 percent. This pre-processing step is crucial for PiCoDeS as it uses lifted feature space to simplify the training scheme while maintaining the information preserved in the Reproducing Kernel Hilbert Space (RKHS). Therefore, the method performance will be severely affected when lifted features are not used.

Each method is trained with the training images to discover the attributes. Then we use the manifold $\mathcal{M}$ w.r.t. the test images for the evaluation. More precisely, each attribute descriptor is extracted from test images (*i.e.,* $\mathbf{z}_k, \mathbf{z}_k \in \{-1, 1\}^N$, where N is the number of test images). For each dataset, we use the attribute labels from AMT to represent the Meaningful Subspace, $\mathcal{S}$.

We adapted the MPPCA metric from the semi-supervised attribute discovery method proposed in [77]. In particular, to discover an attribute, the method in [77] progressively updates MPPCA model using human feedback. In our settings, we directly train MPPCA using attributes found from AMT for each dataset. To measure meaningfulness, we compute the posterior probability of the given discovered attribute to the MPPCA model. We train MPPCA model using five components and three-dimensional subspace for the ASUN dataset. As for the ApAy dataset, we use three components and three-dimensional subspace. This is because the number of attributes in the ApAy dataset is much smaller than the ASUN dataset. Unless otherwise stated, we follow the experiment settings for MPPCA as described in [77]. For instance, we employ a threshold on the posterior probability to determine whether an attribute is meaningful. The MPPCA metric is the calculated by computing the percentage of the attributes deemed as meaningful over the total discovered attributes.

5.4.2 Do δ_{cvx} and δ_{jp} Measure Meaningfulness?

In this experiment, we evaluate whether the proposed approach really does measure the meaningfulness on a set of automatically discovered attributes. One of the key assumptions in our proposal is that the distance between the Meaningful Subspace and the given attribute set $\mathcal{D}$ reflects the meaningfulness of a set of attribute. More specifically, if the distance is small, it is assumed that the attribute set is potentially meaningful and vice-versa. Aiming to evaluate that, we construct two sets of attributes, respectively with meaningful and non-meaningful attributes, and observe their distances to the meaningful subspace.

As to the meaningful attribute set, we follow the methods used in Section 5.3. Providing manually labelled attribute set $\mathcal{S}$, here denoted the AMT attribute set [1], in each dataset, we follow the approach used in Section 5.3 to divide the set into two subsets $\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S}$ where $\mathcal{S}^1$ represents the Meaningful

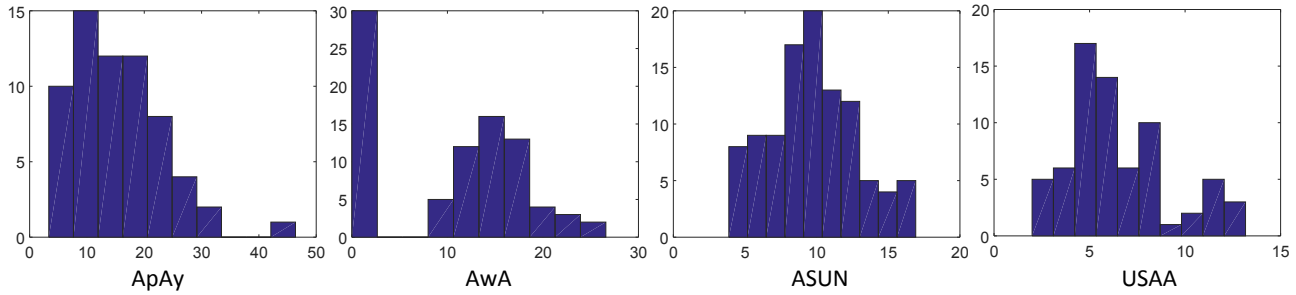
[1] As mentioned before, attributes discovered from the AMT procedure are considered meaningful. For the sake of clarity, we call these attributes AMT attributes.

Subspace and $\mathcal{S}^2$ is considered as a set of discovered attributes (*i.e.,* $\mathcal{D} = \mathcal{S}^2$). Unequivocally, the attributes in $\mathcal{S}^2$ should be meaningful as they are manually labelled by human annotators. Thus, we name $\mathcal{S}^2$ as the *MeaningfulAttributeSet*.

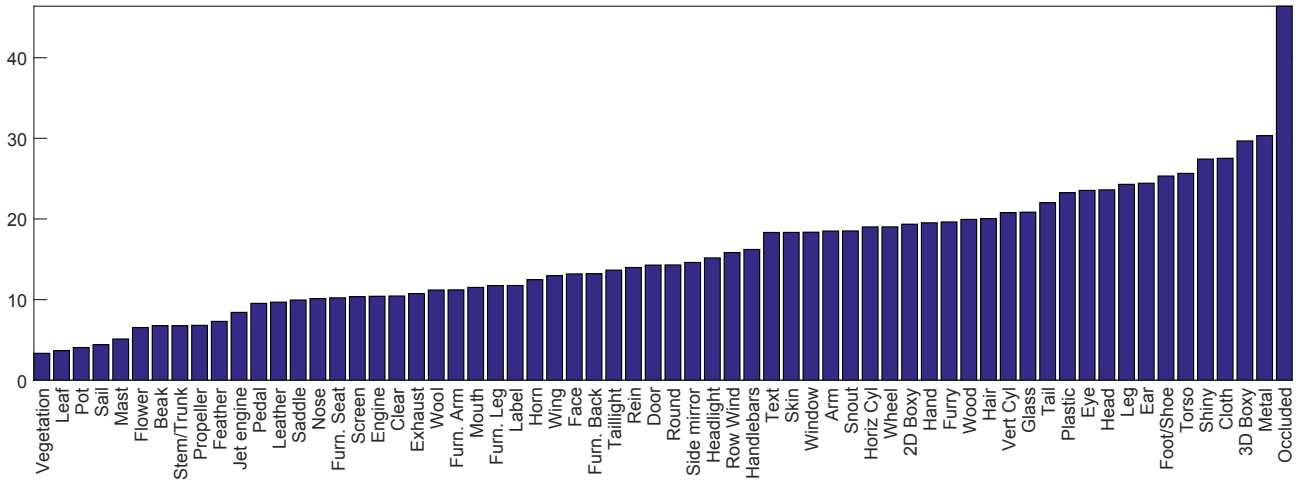
For non-meaningful attribute set, we create this set by randomly generating non-meaningful attributes. As described in Section 5.3, we generate a finite set of random attributes denoted by $\tilde{\mathcal{N}}$. We name this set as *NonMeaningfulAttributeSet* since it is non-meaningful and should have significantly larger distance to the Meaningful Subspace.

Recall that the Meaningful Subspace $\mathcal{S}^1$ needs to be carefully selected to maximise the meaningful subspace spanned. However, to show the efficacy of our proposed selection, we first randomly select $\mathcal{S}^1$. Then, on the second experiment, we apply our proposed selection approach. To perform our proposed selection approach, we must evaluate the independence of each AMT attribute via analysing its attribute name and computing its individual reconstruction error. We will always put independent attributes in $\mathcal{S}^1$. In other words, let $\hat{\mathcal{S}}^1$ be the set of AMT attributes marked as independent attributes. Then, the set $\mathcal{S}$ is divided into $\mathcal{S}^1$ and $\mathcal{S}^2$ such that, $\hat{\mathcal{S}}^1$ will always be in $\mathcal{S}^1$. In this case, we still randomly divide $\mathcal{S}$ with a constraint that $\hat{\mathcal{S}}^1$ should always be in the set $\mathcal{S}^1$. As previously described, a leave-one-attribute-out scheme is used to determine the independence of an AMT attribute with respect to the rest of AMT attribute set. Figure 5.2a presents the result of this analysis. As we can see, the reconstruction errors of ApAy dataset are, in general, much larger than the other datasets. We conjecture that this might be caused by the fact that the other three datasets are all fine-grain classification datasets, however ApAy is proposed for addressing the general classification problem. This means the attributes provided in this dataset are more likely to be independent as they are used to describe a wide variety of classes. For further inspection, we take the ApAy dataset and present the results in the perspective of semantic reasoning of the attribute names. In Figure 5.2b, we plot the reconstruction error for each attribute in ascending order. As we can see, the attributes with low reconstruction errors are often more likely to be correlated (*i.e.,* less independent). For example, images such as the "leaf" and "pot" can reconstruct the "vegetation" attribute, "sail" and "mast" are often present together in the sailing scenes. However, the attributes with high reconstruction error are more likely to be independent (*i.e.,* less correlated) such as shape-related attribute "3D Boxy", material-related attribute "Metal", and especially "Occluded". It is worthy to note that we only show the analysis using reconstruction error with convex hull regularisation. The same findings are also exhibited when the ℓ_0 regularisation is used.

To reasonably determine the parameter α (See Section 5.3.2), we average the highest reconstruction error scores from the other three fine-grain datasets. This gives us a value $\alpha = 18.89$. That means we consider any attributes in the datasets with error above α to be independent attributes. Thus, the 22 independent attributes with highest reconstruction error can be put into the meaningful attribute subset $\mathcal{S}^1$ for better approximation of meaningful attribute subspace. The rest of the attributes are still pooled and randomly selected. Table 5.1 shows the results with, and without, the proposed selection



(a)



(b)

Figure 5.2: The reconstruction error analysis on datasets. In (a), the horizontal axis represents the reconstruction error value; the vertical axis represents the frequency of the attributes which fall into the range of reconstruction error. In (b), the horizontal axis depicts the name of each attribute, the vertical axis represents the reconstruction error value.

strategy. As we can see, after applying the selection, the MeaningfulAttributeSet, which is always considered as meaningful, shows the lowest reconstruction error. That other methods almost remain the same with little random perturbation which indicates that, for the automatic attribute discovery methods, our metric is quite stable. Again, we note that we use δ_{cvx}, and similar results are found when using δ_{jp}.

Table 5.1: Comparisons of reconstruction error results on ApAy dataset with, or without, the selection strategy.

	With selection	Without selection
PiCoDeS	12.65	12.52
DBC	48.97	49.66
ITQ	50.73	51.70
SPH	48.91	49.79
LSH	52.17	53.14
KSH	38.34	38.66
NonMeaningful	53.20	54.34
Meaningful	12.09	18.29

Now, we are ready to discuss the evaluation of our proposed approach to determine whether our approach can measure the attribute meaningfulness. In order to do that, we first perform the subspace interpolation of all the attribute sets discovered by the methods. To perform the subspace interpolation, the random attributes are progressively added to the set of attributes from each method. By doing this, we can evaluate if the distance to Meaningful Subspace is enlarged when we progressively increase the number of non-meaningful attributes.

Figure 5.3 presents the evaluation results.

Here we present the reconstruction error results for all 4 datasets where 16 and 32 attributes are discovered by the methods respectively. Recall that although we carefully selected $\mathcal{S}^1$, the set $\mathcal{S}$ is still randomly divided. We produced the results shown in Figure 5.3 by repeating the random division of $\mathcal{S}^1$ 100 times and calculating the average distance. The detail results are also shown in Table 5.2. We note that no matter how $\mathcal{S}^1$ is selected, our method is relatively stable.

As we can see from the results that the *MeaningfulAttributeSet* has the closest distance to the Meaningful Subspace for both distances δ_{cvx} and δ_{jp} on all datasets. As expected, the *NonMeaningfulAttributeSet* has the largest distance compared with the others. In addition, when the random attributes are progressively added, the distance increases between the Meaningful Subspace and the sets of attributes discovered by each method. These results indicate that the proposed approach could be used to measure the meaningfulness of a set of attributes. Moreover, they also give a strong indication that there is a *shared structure* between meaningful attributes.

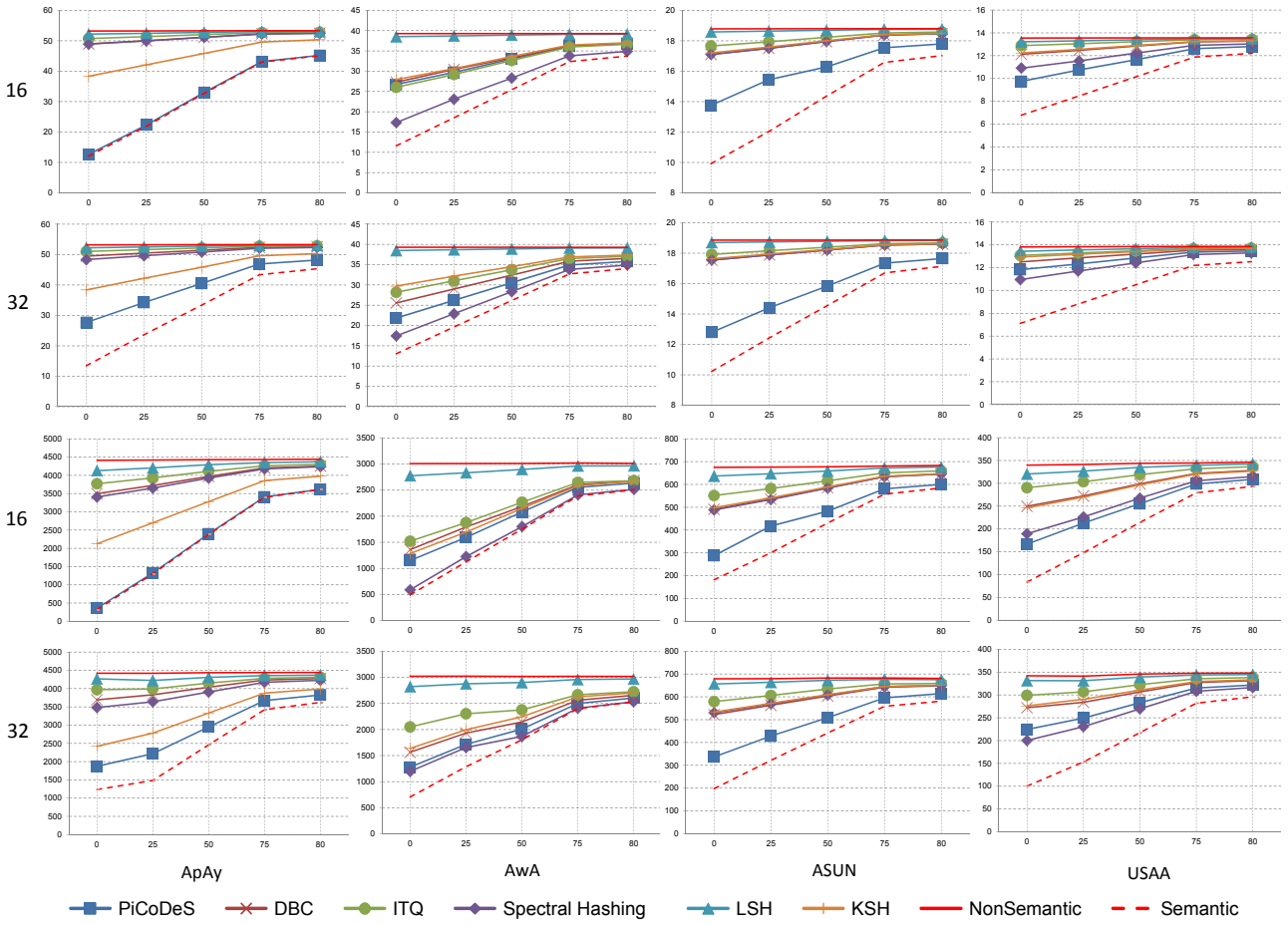


Figure 5.3: Validation of attribute meaningfulness measurement by reconstruction error δ_{cvx} (first and second rows) and δ_{jp} (third and fourth rows). In each subfigure, the horizontal axis represents the percentage of noise attributes and the vertical axis means the reconstruction error values. As we can see, both distances become larger when more random/non-meaningful attributes are added. MeaningfulAttributeSet has the closest distance to the Meaningful Subspace and NonMeaningfulAttributeSet always has the largest distance. Here, each method is configured to discover 16 and 32 attributes. The smaller the δ, the more meaningful the attribute set is.

Table 5.2: The table of values in reconstruction errors with standard deviation analysis. Letter E conventionally represents 'times n raised to the power of 10'

		cvx				jp			
		ApAy	AwA	ASUN	USAA	ApAy	AwA	ASUN	USAA
16	PiCoDeS	12.65 ±9.57E-02	26.73 ±7.10E-02	13.76 ±2.33E-02	9.74 ±5.57E-02	358.61 ±1.53E+01	1150.76 ±1.71E+01	287.95 ±8.80E-01	166.57 ±3.17E+00
	DBC	48.97 ±1.78E-04	27.30 ±2.04E-01	17.10 ±2.19E-02	12.12 ±1.14E-01	3499.92 ±2.93E+00	1358.82 ±2.32E+01	491.13 ±1.01E+00	249.69 ±3.39E+00
	ITQ	50.73 ±3.15E-04	26.04 ±3.51E-01	17.66 ±4.85E-02	12.88 ±6.79E-02	3768.67 ±4.66E+00	1519.80 ±6.30E+01	551.51 ±3.13E+00	290.29 ±1.03E+00
	SPH	48.91 ±1.26E-03	17.32 ±1.12E-01	17.10 ±1.85E-02	10.91 ±3.40E-02	3406.41 ±3.22E+00	588.92 ±3.38E+01	488.34 ±9.30E-01	189.45 ±8.48E-01
	LSH	52.17 ±1.15E-09	38.51 ±7.64E-03	18.58 ±3.91E-02	13.22 ±9.34E-02	4127.82 ±6.09E+00	2775.42 ±3.59E+00	636.73 ±1.35E+00	320.39 ±1.53E+00
	KSH	38.34 ±1.53E-02	27.94 ±9.73E-02	17.21 ±1.91E-02	12.23 ±6.70E-02	2122.98 ±4.50E+00	1285.05 ±9.89E+00	498.27 ±1.48E+00	246.14 ±1.40E+00
	NonMeaningful	53.20 ±1.52E-09	39.27 ±4.32E-14	18.78 ±2.75E-02	13.55 ±1.21E-01	4411.74 ±7.45E-01	3008.71 ±8.16E-01	675.03 ±3.84E-01	339.79 ±2.52E+00
	Meaningful	**12.09 ±1.09E+00**	**11.60 ±2.08E+00**	**9.92 ±6.22E-01**	**6.78 ±6.27E-01**	**322.44 ±9.29E+01**	**492.54 ±1.06E+02**	**182.04 ±2.20E+01**	**83.54 ±1.44E+01**
32	PiCoDeS	27.67 ±1.12E-01	21.84 ±5.42E-02	12.79 ±4.39E-02	11.82 ±1.49E-01	1872.33 ±3.95E+01	1277.40 ±6.75E+01	336.71 ±4.75E+00	224.44 ±1.96E+00
	DBC	49.50 ±1.25E-04	25.56 ±2.69E-01	17.55 ±5.88E-02	12.50 ±3.02E-01	3691.58 ±7.46E+00	1563.89 ±6.90E+01	522.29 ±1.61E+00	272.04 ±2.89E+00
	ITQ	51.02 ±2.57E-04	28.18 ±4.31E-01	17.92 ±1.09E-01	13.07 ±2.09E-01	3971.37 ±1.22E+01	2048.51 ±6.32E+01	578.97 ±3.25E+00	298.99 ±1.66E+00
	SPH	48.35 ±1.62E-03	17.43 ±1.23E-01	17.54 ±4.82E-02	10.95 ±1.08E-01	3480.28 ±2.39E+01	1196.88 ±1.01E+02	530.47 ±1.45E+00	200.20 ±2.80E+00
	LSH	52.22 ±3.37E-09	38.48 ±1.12E-02	18.70 ±9.51E-02	13.43 ±2.99E-01	4268.15 ±9.82E+00	2822.03 ±1.02E+01	656.52 ±2.36E+00	331.45 ±1.54E+00
	KSH	38.37 ±3.52E-02	29.71 ±1.29E-01	17.63 ±4.34E-02	12.93 ±2.68E-01	2419.02 ±1.52E+01	1637.47 ±5.03E+01	533.15 ±1.46E+00	276.13 ±1.95E+00
	NonMeaningful	53.19 ±6.56E-10	39.28 ±1.18E-04	18.86 ±8.02E-02	13.82 ±3.73E-01	4421.49 ±1.13E+00	3020.82 ±1.69E+00	678.89 ±6.81E-01	342.10 ±1.47E+00
	Meaningful	**13.45 ±5.79E-01**	**13.04 ±1.51E+00**	**10.23 ±5.13E-01**	**7.12 ±4.13E-01**	**1234.20 ±9.14E+01**	**706.19 ±9.04E+01**	**197.50 ±1.50E+01**	**100.29 ±7.55E+00**

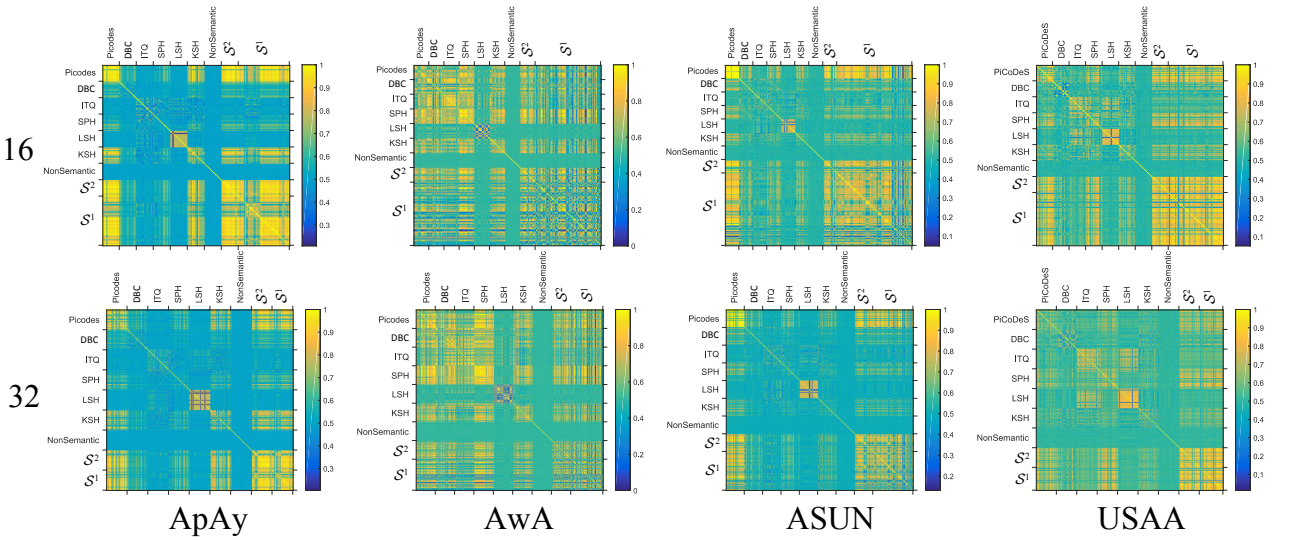


Figure 5.4: Visualisation of co-occurrence matrix, the colourbar represents the value of joint probability. The range between two ticks represents the attributes from each method. The first row shows the result for each dataset when each method is configured to discover 16 attributes. The second row shows the result for each dataset when each method is configured to discover 32 attributes.

5.4.3 Attribute Co-occurrence Matrix Analysis

For further inspection, we also perform the co-occurrence matrix analysis on the attributes discovered by each method and the AMT attributes *i.e.,* $\mathcal{S}^1$ and $\mathcal{S}^2$. The results are shown in Figure 5.4.

The co-occurrence matrix figures represent the visualisation of joint probability between the discovered attributes from each method and the AMT attributes, which are considered as meaningful. As we can see in the figure, in almost every dataset, the highest joint probability is achieved between $\mathcal{S}^1$ and $\mathcal{S}^2$. The trend is obvious in ApAy dataset, ASUN dataset and USAA dataset. However, the trend does not look apparent in AwA dataset. We conjecture this could be that many attributes in $\mathcal{S}$ are independent. The attributes in AwA dataset are class-level *i.e.,* each sample in same class has the same attribute representation. Therefore, in order to guarantee the discriminative power between classes, the attributes may be chosen to reflect different aspects of classes, thus they could have lower joint probability. We note that the supervised attribute learning methods such as PiCoDeS, DBC, and KSH also have comparable high probability with the AMT attributes. An interesting finding is that the attribute representations generated by the LSH method tend to have high joint probability with each other. This may be due to the simple linear projection of the data feature matrix in generating the final attribute representation. Generally, the trend of the results is consistent with the previous experiments, which further indicates the capability of our approach to capture the attribute meaningfulness.

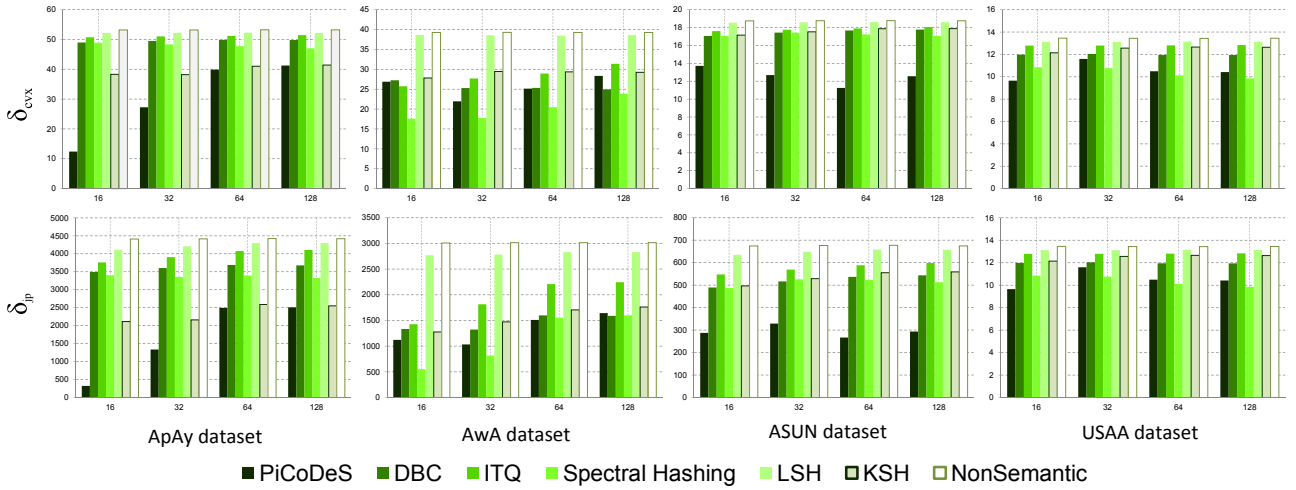


Figure 5.5: *Attribute meaningfulness* comparisons between different methods on variant number of discovered attributes. The first row reports the results using δ_{cvx} and the second row reports the results using δ_{jp}. The smaller the δ, the more meaningfulness.

5.4.4 Attribute Set Meaningfulness Evaluation Using δ_{cvx} and δ_{jp}

In this section, the meaningfulness is evaluated by δ_{cvx} and δ_{jp} for the set of attributes automatically discovered by various comparative methods. For that purpose, all manually labelled attributes from AMT in each dataset are used as the representation of the Meaningful Subspace. Then each method is configured to discover 16, 32, 64, and 128 attributes.

Figure 5.5 reports the evaluation results on all datasets. It is noteworthy to mention that both the proposed distances δ_{cvx} and δ_{jp} are not calibrated and scaled, making it difficult to perform in-depth evaluation. However, we still can evaluate the results in terms of the method rank ordering (*i.e.,* which method takes first place and which comes second).

PiCoDeS has the smallest distance in various numbers of attributes extracted on most of the datasets. PiCoDeS applies category labels and max-margin framework to jointly learn the category classifier, and attribute descriptor in an attempt to maximise the discriminative power of the descriptor. In other words, the goal of PiCoDeS is to discover a set of attributes as sample representations which can discriminate between categories.

DBC is also developed under the maximum-margin framework to extract meaningful attributes as PiCoDeS. However, compared with PiCoDeS, DBC discovers less meaningful attributes. We conjecture the reason could be DBC learns the whole attribute descriptor for each category simultaneously, unlike PiCoDeS that learns the attribute individually. This scheme will inevitably emphasise category discrimination of attributes rather than preserving the meaningfulness of individual attribute. Note that here we do not suggest that DBC is not able to discover meaningful attributes, rather, PiCoDeS may find more meaningful attributes. Therefore, our finding does not contradict the results presented in the DBC original paper [86].

Another observation from the results of SPH indicates that it is able to discover meaningful attributes. SPH is aimed to discover binary codes via a graph embedding approach, preserving the local neighbourhood structure. One possible explanation could be that when two images belong to the same class, they should share more attributes, indicating a shorter distance between them in the binary space, and vice-versa.

Although ITQ aims to learn similarity preserving binary descriptor, it has a larger distance than SPH, DBC, and PiCoDeS. The reason may be the way ITQ learns the binary descriptor, which mainly relies on the global information of the data distribution. In other words, the algorithm minimises the quantisation error of the mapping data to the vertices of a zero centred binary hypercube, suggesting that only global information by itself might not be sufficient to discover meaningful attributes.

As expected, the attribute sets from LSH have the largest distances to the Meaningful Subspace (*i.e.,* least meaningfulness). LSH uses random hyperplanes to project a data point into the binary space. Therefore, the consistent identifiable visual concepts are hardly presented in the positive images.

In summary, two recipes could be derived from the current results that could be significant for future automatic attribute discovery method design: the method should attempt to preserve local neighbourhood structure, as well as to consider the discriminative power of attributes.

5.4.5 Attribute Set Meaningfulness Calibration Using the Proposed Meaningfulness Metric

As described in Section 5.4.4, the distance between attribute sets and the meaningful subspace has some limitations preventing us from performing an in-depth analysis. Quantitative comparisons between different methods are more desirable in the analysis of attribute meaningfulness. Here we show the meaningfulness metric.

We apply γ_{cvx} and γ_{jp} on the datasets and for each method through calibrating the proposed distances δ_{cvx} and δ_{jp}.

The results are shown in Figure 5.7 when each method is configured to discover 16 and 32 attributes. The rank orders of the results of the methods are almost the same with similar values for metric γ_{cvx} and γ_{jp}. There are only two exceptions in ASUN dataset. This can be explained by the fact that each metric captures a different aspect of attribute meaningfulness. The proposed γ_{cvx} captures a one-to-many relationship while γ_{jp} evaluates the one-to-one relationship. Then the equal weighted metric score $\tilde{\gamma}$ is applied for further analysis.

A user study is also conducted on the attributes discovered by each method. Since AwA requires experts in animal studies and USAA is a large video dataset whose complex social group activities are likely to cause subjective bias, we only use ApAy and ASUN datasets for the user study.

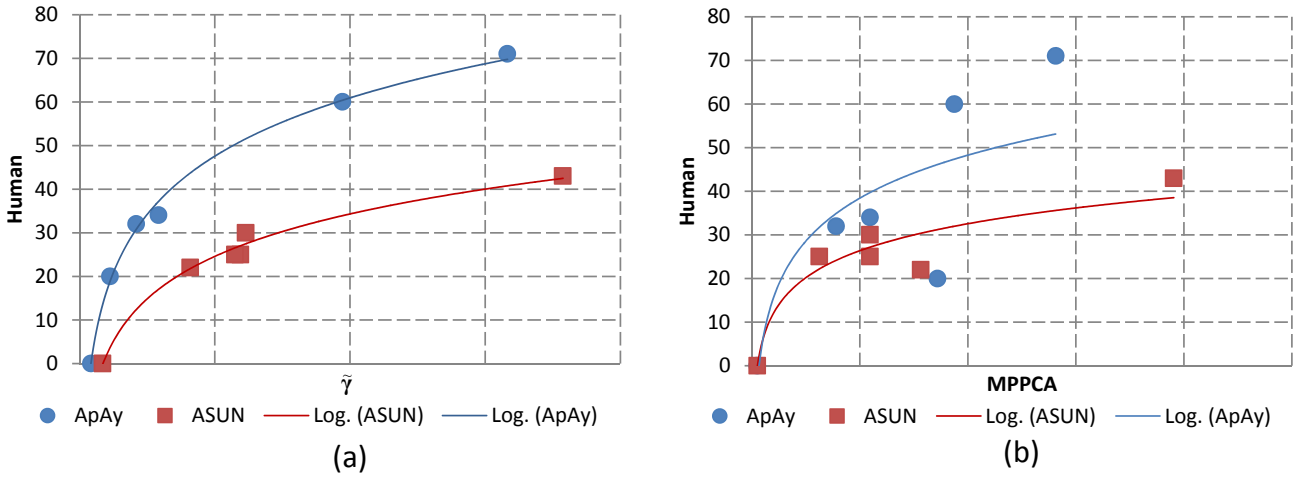


Figure 5.6: Demonstration of correlation analysis between user study and the proposed method $\tilde{\gamma}$ as well as MPPCA on both ApAy and ASUN datasets.

The study collected over 100 responses for each number of discovered attributes. In each response, there are positive and negative images presented from 8 randomly chosen discovered attributes. The user was asked whether these two set of images represent a consistent visual concept (hence meaningful). The users are university staff and students with different knowledge background from various major fields including IT, Electronic Engineering, History, Philosophy, Religion and Classics, and Chemical Engineering. The responses were averaged by considering 1 as meaningful and 0 as non-meaningful.

Table 5.3 illustrates the result of $\tilde{\gamma}$ compared with the human study. Again, the attribute set discovered by LSH has the lowest meaningful content at close to 0%. Thus, LSH generates the least meaningful attribute sets. PiCoDeS and SPH generally discover more meaningful attribute sets. The methods using a randomisation scheme such as LSH and ITQ tend to generate less meaningful attribute sets with attribute meaningfulness around 1%-20%. The results indicate that the attribute meaningfulness could be significantly increased (*i.e.,* on average by 10-20 percentage points) by applying learning techniques such as PiCoDeS, DBC, and SPH.

Compared with the results of the proposed metric $\tilde{\gamma}$, similar trends have been observed in the user study. Moreover, the user study results compared with γ_{cvx} and γ_{jp} are also shown in Figure 5.7. Consistent similar trend as shown in previous experiments is visible.

As for MPPCA metric, similar results can be found, such as LSH discovers the least meaningful attribute sets, and PiCoDeS generally discovers more meaningful attribute sets. However, the result of our proposed method is closer to the human study in terms of the ranking order of attribute discovery methods. This could indicate that the amount of AMT attributes used to train the MPPCA may not be sufficient. We note that, the MPPCA was originally designed to have human feedback in multiple

iterative processes to discover attributes [77]. Since in our experiment we only fed the MPPCA once with the AMT attributes, it may lack human feedback.

We also perform two statistical analyses to compare which metric is closer to the human study. Both of the analyses are shown in Figure 5.6 by applying a simple logarithmic fitting using the data from Table 5.3. Figure 5.6(a) shows that regression line fits these data very well. The coefficients of determination R^2 [25], which indicates how well data fit a statistical model, for ApAy dataset and ASUN dataset fitting are respectively 0.99 and 0.98 between the proposed metric and the human study results. The results suggest the regression line of our proposed method nearly perfectly fits the data. Figure 5.6(b) shows the fitting result for MPPCA is not as good as the proposed method. The R^2 values of MPPCA metric are respectively 0.64 and 0.89 on ApAy dataset and ASUN dataset.

This demonstration further indicates to some extent, our proposed metric is able to evaluate the meaningfulness of a set of discovered attributes from comparative methods, similar to how a human does, via a simple non-linear regression.

It is noteworthy to mention that the time cost of the evaluation by our metric is significantly lower than the manual process using AMT. Recall that, the time required for a human annotator (an AMT worker) to finish one HIT is 2 minutes, an AMT worker may need 320 minutes to finish evaluating 5 methods wherein each is configured to discover 32 attributes. Our approach only needs 105 seconds in total to evaluate all four datasets (*i.e.,* average 35 seconds each complete dataset); thus, leading to several orders of magnitude speedup!

Table 5.3: The results (in percentage) of meaningfulness metric $\tilde{\gamma}$ on each dataset compared with user study and MPPCA metric on ApAy & ASUN datasets. Each method is configured to discover 32 attributes. In addition, for convenience we also report the proposed metric results on AwA & USAA datasets. The bold text indicates the top performing method in the proposed metric. The higher, the more meaningful.

Methods \Datasets	ApAy			ASUN			AwA		USAA	
	$\tilde{\gamma}$	MPPCA	Human	$\tilde{\gamma}$	MPPCA	Human	$\tilde{\gamma}$	Human	$\tilde{\gamma}$	Human
LSH	1.7	0	0	3.4	0	0	5.6	N/A	4.7	N/A
ITQ	4.5	34.4	20	16.4	31.3	22	41.6		13.1	
SPH	11.7	21.9	34	23.8	21.9	25	**80.7**		**48.3**	
DBC	8.4	15.6	32	24.6	21.9	30	57.1		22.3	
KSH	38.9	37.5	60	23.0	12.5	25	47.6		18.3	
PiCoDeS	**63.3**	**56.3**	**71.0**	**71.5**	**78.1**	**43**	70.5		36.8	

5.5 Summary

In this chapter, we studied a novel problem of measuring the meaningfulness of automatically discovered attribute sets. To that end, we proposed a novel metric, here called the *attribute meaningfulness* metric. We developed two distance functions for measuring the meaningfulness of a set of attributes. The distances were then calibrated by using subspace interpolation between Meaningful Subspace and Non-meaningful/Noise Subspace. We proposed the meaningful attribute set selection technique that led to a better meaningful subspace approximation. The final metric score indicated how much meaningful content was contained within the set of discovered attributes. In the extensive experiment, the proposed metrics were used to evaluate the *meaningfulness* of attributes discovered by two recent automatic attribute discovery methods and four hashing methods on four datasets. A user study on two datasets showed that the proposed metric has strong correlation to human responses. Our metric was also shown to be more correlated with the user study compared with a metric adapted from a recent semi-supervised attribute discovery method. All results suggested that there is a strong indication that the shared structure may exist among the meaningful attributes. The results also suggest that discovering attributes by optimising the attribute descriptor discrimination and/or preserving the local similarity structure could yield more meaningful attributes. In future work, we plan to explore other constraints or optimisation models to capture the hierarchical structure of semantic concepts. We also plan to perform more large-scale user studies using AMT on other datasets.

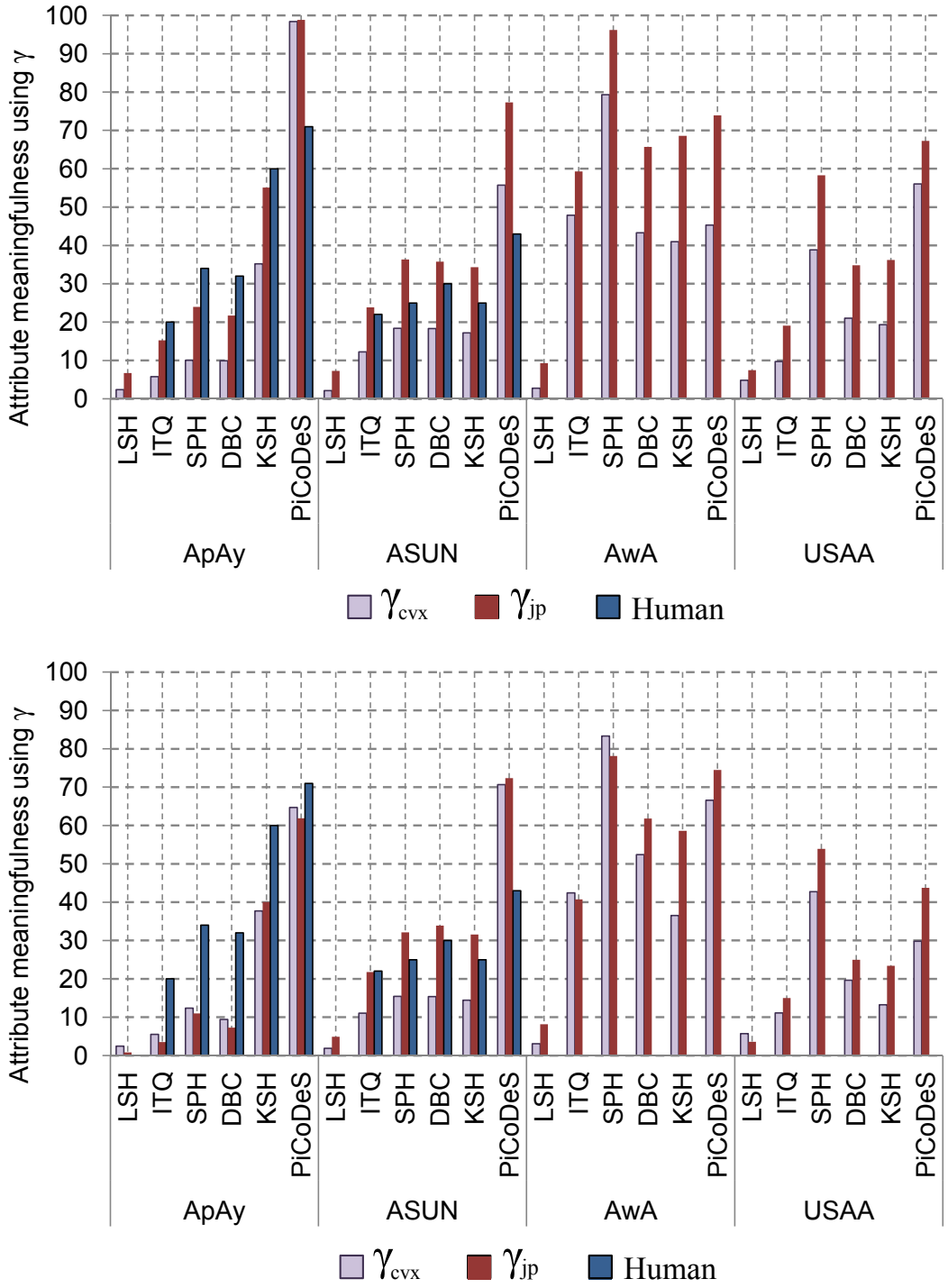


Figure 5.7: Comparisons of various methods using the proposed meaningfulness metric as well as human study results. Each method is set to discover 16 and 32 attributes. The higher, the more meaningful. Human study is not conducted for AWA dataset, as special zoology knowledge is required, nor for USAA dataset due to inconvenience to display and subjectiveness bias problem. The human results for LSH method are 0 for ApAy and ASUN datasets.

Chapter 6

Determining the Best Attributes for Surveillance Video keyword generation

> There is no doubt that our nation's security and defeating terrorism trump all other priorities.
>
> Arlen Specter

Chapter Summary: *In this chapter, we introduce an application study of attribute meaningfulness evaluation method for surveillance video keyword generation. We propose a novel approach, based on the shared structure exhibited amongst meaningful attributes, that enables us to compare between different automatic attribute discovery approaches. We then validate our approach by comparing various attribute discovery methods on two attribute datasets.*

6.1 Introduction

In this chapter, we introduce the application study of the automatic evaluation method of visual attribute proposed in Chapter 5. We show that the evaluation method is quite useful for determining that the automatic attribute discovery method can extract the most meaningful keywords for surveillance videos. Thus, this application study demonstrates that visual attribute is quite a useful tool which especially helps the automatic video analysis. This answers the research question 1 in Section 1.1 furthermore, from the application study point of view.

Automatic video analytics is one of the key components in smart surveillance systems to combat crime and terrorism. For example, they can be used to detect anomalous events to alert security officers [44]. In general, surveillance systems generate a large amount of video data. This makes finding critical information in the surveillance video as challenging as finding the proverbial needle

in a haystack [83]. Thus automation is highly desirable so one can reduce the amount of time to find this critical information.

Automatic video analytics have been gaining significant interest in the research community. Some examples of the current works are: action recognition [93], face hallucination [59], anomaly detection [70], and video description [72].

In this chapter, we tackle the problem of automatic generation of keywords for video description. Keywords are important ingredients in generating textual descriptions [90]. More specifically, once the keywords of a video are generated, the video can be searched using natural language to find events of interest.

Unfortunately, existing approaches still require a great deal of manual labelling before the systems can be used to generate the keywords/description [37]. For example, the work proposed by Izadinia and Shah in [42] uses extensive spatio-temporal annotations to train action and role models for action recognition. The approach produces better descriptions than many other approaches. However, the significant manual labelling severely restricts its scalability. In addition, when relevant manual labels are not available, then it is not possible for the system to describe unusual events which would be extremely useful in anomaly event analysis.

One feasible way to circumvent this is to employ latent hierarchical probabilistic models such as probabilistic Latent Semantic Analysis (pLSA) [104] or Latent Dirichlet Allocation (LDA) [111]. These methods can automatically mine the latent topics which could represent keywords. Thus, when a topic is inferred in a video, then the associated text of the topic becomes the keyword. Unfortunately, despite their potential, these methods are based on the bag-of-words model requiring explicit modelling of visual words. Here, each video is assumed to have a collection of visual words. This explicit assumption may not be feasible for other recent video features not derived from the bag-of-words features.

Inspired by the probabilistic latent topic discovery methods, in this chapter, we propose a method that can automatically discover video keywords with significantly less manual processing. More specifically, several attribute discovery methods such as PiCoDeS [4] and Spectral Hashing [113] can be employed.

Visual attribute features are binary features indicating the presence/absence of visual concepts. For instance, a car can be described as ['has wheels', 'is metallic', 'does not have legs']. In practice, we can represent the binary features as [1 1 0]. The attribute features trained in one domain can be reused for another domain with minimum manual work [28]. As such, a system can be potentially trained to recognise unseen events [51].

Visual attributes have shown promising results in many works which deal with video related tasks [90, 85] as well as in some novel problems such as the zero-shot learning problem [51].

Once the attribute features are trained, they can be used to extract keywords. Unfortunately, training attribute feature also requires extensive manual labelling work. This is because, as each

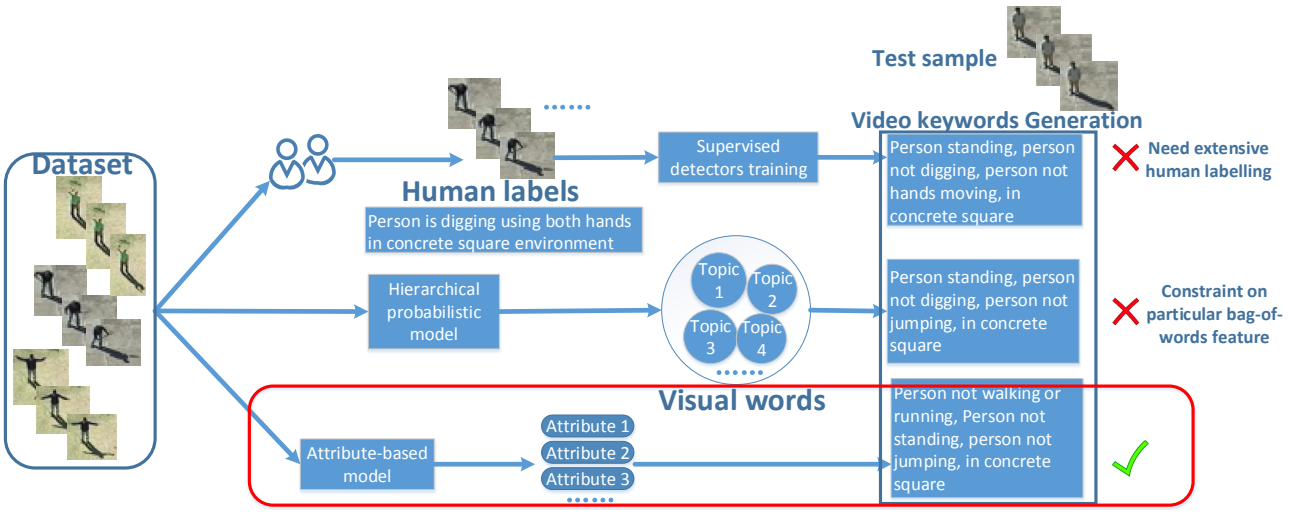


Figure 6.1: The comparisons and properties between our approach and other existing methods on video keyword generation.

individual visual attribute is a binary classifier, then one needs to create the labelled training set for each attribute.

To that end, some researchers have turned their attention to automatic attribute discovery methods [4, 86, 115]. These methods primarily focus on learning an embedding function that maps the original descriptors into binary code space wherein each individual bit is expected to represent a visual attribute. We note that these approaches are also closely related to hashing approaches [34, 54, 113]. The difference is that, unlike automatic attribute discovery approaches, hashing methods are primarily aimed at significantly reducing computational complexity and storage whilst maintaining system accuracy. Despite many works that have been proposed, it is not clear which methods produce the most meaningful attributes.

Here, we present an approach that allows us to select the attribute discovering method that discovers the most meaningful attributes. We then find the keywords extracted from the best method to describe videos recorded from a surveillance system.

The intuition of our approach comes from a speculation proposed in [76, 77]. More specifically, Parikh and Grauman suggest that meaningful attributes tend to occupy a subspace, here called the Meaningful Subspace, on a manifold. Thus, we can utilise any given set of meaningful attributes to be our 'yardstick' for comparing various attribute discovery methods.

Figure 6.1 illustrates the differences between our attribute-based keyword generation approach and the existing approaches. We can see our approach has two main advantages. First, is it does not require significant manual processing. Second, it is not constrained to one particular video feature.

Contributions: We list our contributions as follows: (1) We propose an attribute-based video keyword generation approach by utilising the attribute discovery method that discovers the most meaningful set of attributes; (2) To determine the attribute discovery method, we propose a selection approach

enabling us to select which attribute discovery methods provide meaningful attributes; (3) We use and validate our selection method in two known attribute datasets; (4) Finally, we validate the keywords extracted from the best attribute discovery method. These keywords can be used to describe videos recorded from a surveillance system.

We continue this chapter as follows. Section 6.2 presents our proposed approach to compare various attribute discovery methods. Section 6.3 describes the approach to generate video keywords using discovered attributes. Section 6.4 presents experiments, and Section 6.5 concludes the discussion.

6.2 Selecting the Attribute Discovery Method

As we have introduced in Section 5.2.1, the manifold of decision boundaries of visual attribute w.r.t $\mathcal{X}$ is defined as $\mathcal{M}^{[\mathcal{X}]} \in \{-1, +1\}^N$ which is embedded in an N-dimensional binary space.

As discussed in Section 5.2.2, the previously human labelled attributes from AMT can be considered as a meaningful attribute subset. We define this set of meaningful attributes as $\mathcal{S} = \{\mathbf{h}_j\}_{j=1}^{J}, \mathbf{h}_j \in \{-1, +1\}^N$. The distance between the discovered attribute set and meaningful attribute subset is able to describe the meaningfulness of the discovered attributes. As shown in (5.3), we employ the convex hull regularisation which has been shown in [10] to induce sparsity, avoiding the dense reconstruction coefficients that emerged in (5.1) and (5.2).

The optimisation problem in (5.3) could be solved using the method proposed in [10]. In our approach, we assume that the lower the distance of a set of discovered attributes to a meaningful subspace, the more meaningful the attributes will be.

6.3 Generating Keywords using Discovered Attributes

Once meaningful attributes are discovered, one can extract the attribute features from the given data. However, one still needs to name the attributes. Despite this manual process, we argue that the manual process for naming meaningful attributes is significantly easier and quicker than the manual process of labelling images/videos to train attribute features.

One can name an attribute by first extracting the attribute features from a given set of images. As previously mentioned, each attribute divides any set of images/videos into two groups: the group of images in which the visual attribute is present (the positive class) and the group of images/videos in which the visual attribute is absent (the negative class).

Some attributes may have similar names. In this case, these attributes are considered as duplicate and therefore they are merged.

6.4 Experiment

In this section, we validate our proposed approach and evaluate the accuracy of the keywords extracted from the best discovered method to describe videos.

In the first part, we evaluate the ability of our approach to measure the meaningfulness of a set of attributes. Then, we use our proposed approach to evaluate attribute meaningfulness on the attribute sets generated from various automatic attribute discovery methods such as PiCoDeS [4], as well as the hashing methods such as Spectral Hashing (SPH) [113] and Locality Sensitivity Hashing (LSH) [54]. For this case, two datasets will be utilised: (1) a-Pascal a-Yahoo dataset (ApAy) [28]; (2) SUN Attribute dataset (ASUN) [80].

In the second part of our experiment, we apply the best attribute discovery method to discover keywords from a surveillance dataset. In this setting, we utilise the UT Tower aerial view dataset (UTTower) [16]. The efficacy of the keywords is then evaluated.

6.4.1 Datasets and Experiment Setup

The following is the detailed description of each image dataset for validating our approach and evaluating the attribute discovery methods.

a-Pascal a-Yahoo dataset (ApAy) [28] — comprises two sources: a-Pascal and a-Yahoo. There are 12,695 cropped images in a-Pascal that are divided into 6,340 for training and 6,355 for testing with 20 categories. The a-Yahoo set has 12 categories disjoint from the a-Pascal categories. Moreover, it only has 2,644 test exemplars. There are 64 attributes provided for each cropped image. In total the dataset has 15,339 exemplars, 64 attributes and 32 categories. The dataset provides four features for each exemplar: local texture; HOG; edge and colour descriptor. These are then concatenated into a 9,751 dimensional feature vector. We use the training set for discovering attributes and we perform our study in the test set. More precisely, we consider the test set as the set of images $\mathcal{X}$.

SUN Attribute dataset (ASUN) [80] — ASUN is a fine-grained scene classification dataset consisting of 717 categories (20 images per category) and 14,340 images in total with 102 attributes. There are four types of features provided in this dataset: (1) GIST; (2) HOG; (3) self-similarity; and (4) geometric context colour histograms (See [119] for feature and kernel details). From 717 categories, we randomly select 144 categories for discovering attributes. As for our evaluation, we randomly select 1,434 images (*i.e.,* 10% of 14,340 images) from the dataset. It means, in our evaluation, some images may or may not come from the 144 categories used for discovering attributes.

For the first experiment, we apply the following pre-processing described in [4]. We first lift each feature into a higher-dimensional space approximating the histogram intersection kernel by using the explicit feature maps proposed by Vedaldi and Zisserman [105]. More precisely, each feature is mapped into the space three times larger than the original space. This effectively allows us to apply

linear classifiers in the explicit kernel space [4]. After the features are lifted, we then apply PCA to reduce the dimensionality of the feature space by 40 percent. This pre-processing step is crucial for PiCoDeS as it uses lifted feature space to simplify their training scheme while maintaining the information preserved in the Reproducing Kernel Hilbert Space (RKHS). Therefore, the method performance will be severely affected when lifting features are not used. In our empirical observations, we also found that lifted feature space gives positive contributions to the other methods.

Each method is trained using the training images to discover the attributes. Then we use the manifold $\mathcal{M}$ w.r.t. the test images for the evaluation. More precisely, each attribute descriptor is extracted from test images (*i.e.,* $\mathbf{z}_k, \mathbf{z}_k \in \{-1, 1\}^N$, where N is the number of test images). For each dataset, we use the attribute labels from Amazon Mechanical Turk (AMT) to represent the Meaningful Subspace, $\mathcal{S}$.

UT Tower aerial view activity classification dataset (UTTower) [16] — consists of 108 low-resolution video sequences from 9 types of actions. Each action is performed 12 times by 6 individuals. The dataset is composed of two types of scenes: concrete square and lawn. There are 4 actions in the concrete square scene, they are "pointing", "standing", "digging", "walking", and 5 actions in the lawn scene: "carrying", "running", "wave1", "wave2", "jumping". Ground truth labels for all actions videos are provided for the training and the testing.

For the second experiment, we use the manifold feature proposed in [126] to extract visual information from the surveillance videos in the dataset. The video frames were first downsized into 16 $\times$ 16 pixels and then Grassmann points on $\mathcal{G}_{128,8}$ were generated by performing the Singular Value Decomposition (SVD) on the normalised pixel intensities of 8 successive frames. In total, there are 216 manifold points. Note that, the features are not derived from the bag-of-words framework. It is also noteworthy to mention that our work is not primarily aimed to study feature discriminative power and robustness. Although, it is generally assumed that better features may provide more meaningful attributes, further studies are required in the future.

6.4.2 Attribute Meaningfulness Evaluation

In this experiment, our aim is to verify whether the proposed approach does measure meaningfulness on the set of discovered attributes. One of the key assumptions in our proposal is that the meaningfulness is reflected in the distance between the meaningful subspace and the given attribute set, $\mathcal{D}$. That is, if the distance is far, then it is assumed that the attribute set is less meaningful, and vice-versa. In order to evaluate this assumption we create two sets of attributes, meaningful and non-meaningful attributes, and observe their distances to the meaningful subspace.

For the meaningful attribute set, we use the attributes from AMT provided in each dataset. More precisely, given manually labelled attribute set $\mathcal{S}$, we divide the set into two subsets $\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S}$. Following the method used in Section 6.2, we use $\mathcal{S}^1$ to represent the Meaningful Subspace and

consider $\mathcal{S}^2$ as a set of discovered attributes (*i.e.,* $\mathcal{D} = \mathcal{S}^2$). As human annotators are used to discover $\mathcal{S}^2$, these attributes are considered to be meaningful. We name this as the *MeaningfulAttributeSet*.

For the latter, we create attributes that are not meaningful by random generation. Note that random generation is important to ensure the division is not subjective. More precisely, we generate a finite set of random attributes $\tilde{\mathcal{N}}$. As the set $\tilde{\mathcal{N}}$ is non-meaningful, it should have a significantly large distance to the Meaningful Subspace. We name this set as *NonMeaningfulAttributeSet*. Furthermore, we progressively add random attributes to the set of attributes discovered from each method, to evaluate whether the distance to Meaningful Subspace is enlarged when the number of non-meaningful attributes increases.

Figure 6.2 presents the evaluation results where the methods are configured to discover 32 attributes. From the results, it is clear that *MeaningfulAttributeSet* has the closest distance to the Meaningful Subspace in all datasets. As expected the *NonMeaningfulAttributeSet* has the largest distance compared with the others. In addition, as more random attributes are added, the distance between the sets of attributes discovered for every approach and the Meaningful Subspace increases. These results indicate that the proposed approach could measure the set of attribute meaningfulness. In addition, these also give a strong indication that meaningful attributes have the *shared structure*.

The results presented in Figure 6.2 suggest that PiCoDeS consistently discovers the most meaningful attributes on both datasets. SPH is the second best method to discover meaningful attributes. PiCoDeS utilises max-margin framework to discover the attributes whereas SPH uses spectral relaxation to preserve the similarity between data points in the binary space. In addition, as expected LSH, employing random projection approach, is one of the worst performing methods.

6.4.3 Generating Video Keywords using Discovered Attributes

In this experiment, we will follow the strategy proposed in Section 6.3. Here we ask experts to perform the attribute naming task for the three attribute discovery methods such as PiCoDeS, SPH and LSH configured to discover 16 attributes on the UTTower surveillance video dataset. In this task, we ask 3 experts in surveillance system. The voting mechanism in this task is that when 2 of the 3 experts can name the attribute and the results are similar in semantics, we consider it as the nameable attribute. Then we will use the named attributes as the keywords. To make our work reproducible, our experiment results will be available online[1].

Note that we only take into account the attributes that can be named by experts. This means, any attribute that cannot be named will not be considered as a valid keyword. After performing this task, we found that there are 9 attributes for PiCoDeS, 8 attributes for SPH, and 3 attributes for LSH that can be named. These results suggest that our proposed approach is capable of guiding us in selecting

[1] `http://www.itee.uq.edu.au/sas/datasets`

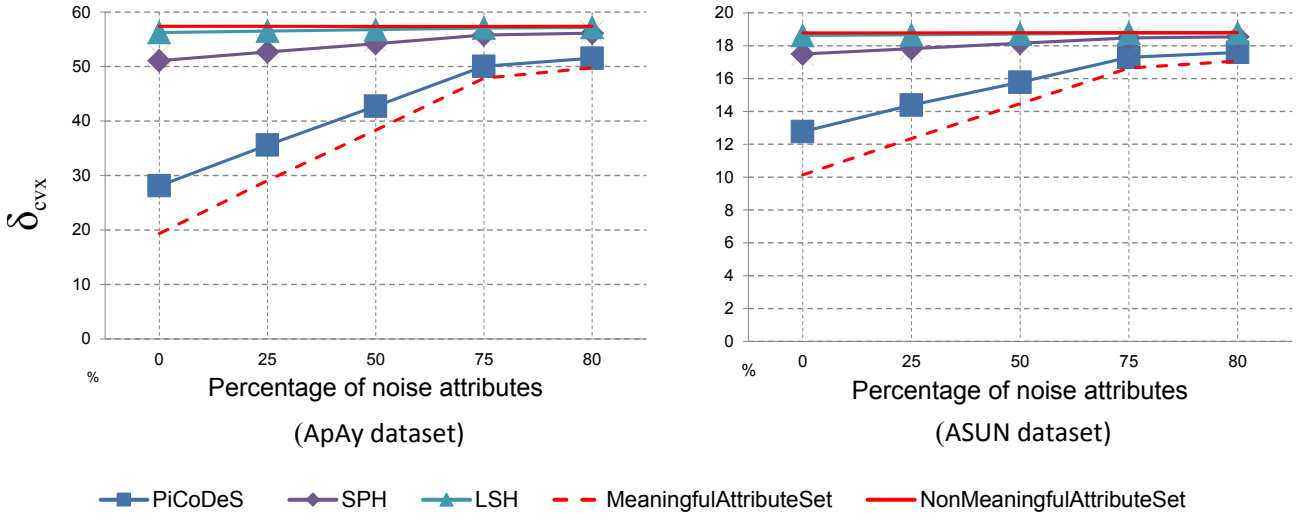


Figure 6.2: Validation of attribute meaningfulness measurement by reconstruction error δ_{cvx}. As we can see, the distances become larger when more random/non-meaningful attributes are added. MeaningfulAttributeSet has the closest distance to the Meaningful Subspace and NonMeaningfulAttributeSet always has the largest distance. Here, each method is configured to discover 32 attributes. The smaller the δ, the more meaningfulness.

the best attribute discovery methods as the experts are able to name most of the discovered attributes by PiCoDeS and SPH.

Once attributes are named, the next step is to generate keywords of each video. Technically, the attributes are extracted from each video. Then, the keywords are generated using the terms of the associated positive attributes.

We evaluate the quality of the generated keywords to describe each video. We then ask human experts to determine whether a keyword is suitable to describe a video.

Figure 6.3 presents two examples where videos are described with suitable keywords, and two examples where videos are described with unsuitable keywords. The examples depicted in Figure 6.3, (a), (b), (c), and (d) are videos of digging, standing, carrying, and waving, respectively.

We count the number of keywords correctly used in each video description and compute the correct hit rates for the whole testing set. The correct hit rate for PiCoDeS, SPH, and LSH are 77.7%, 55.9%, and 48.3%, respectively. This further validates our proposed approach to measure attribute meaningfulness. In addition, it also shows that using the best attribute discovery method, we can automatically generate keywords for videos in a more economical way. Figure 6.4 presents further results in this evaluation. In particular, (a) and (b) report the hit rate for PiCoDeS of each attribute and action, respectively. The plots in (c) and (d) are the hit rate for SPH of each attribute and action, respectively. Most attributes discovered by PiCoDeS have more than 70% hit rate with two attributes having 100% hit rate (all correct). The hit rate for each action also demonstrates an overall good hit

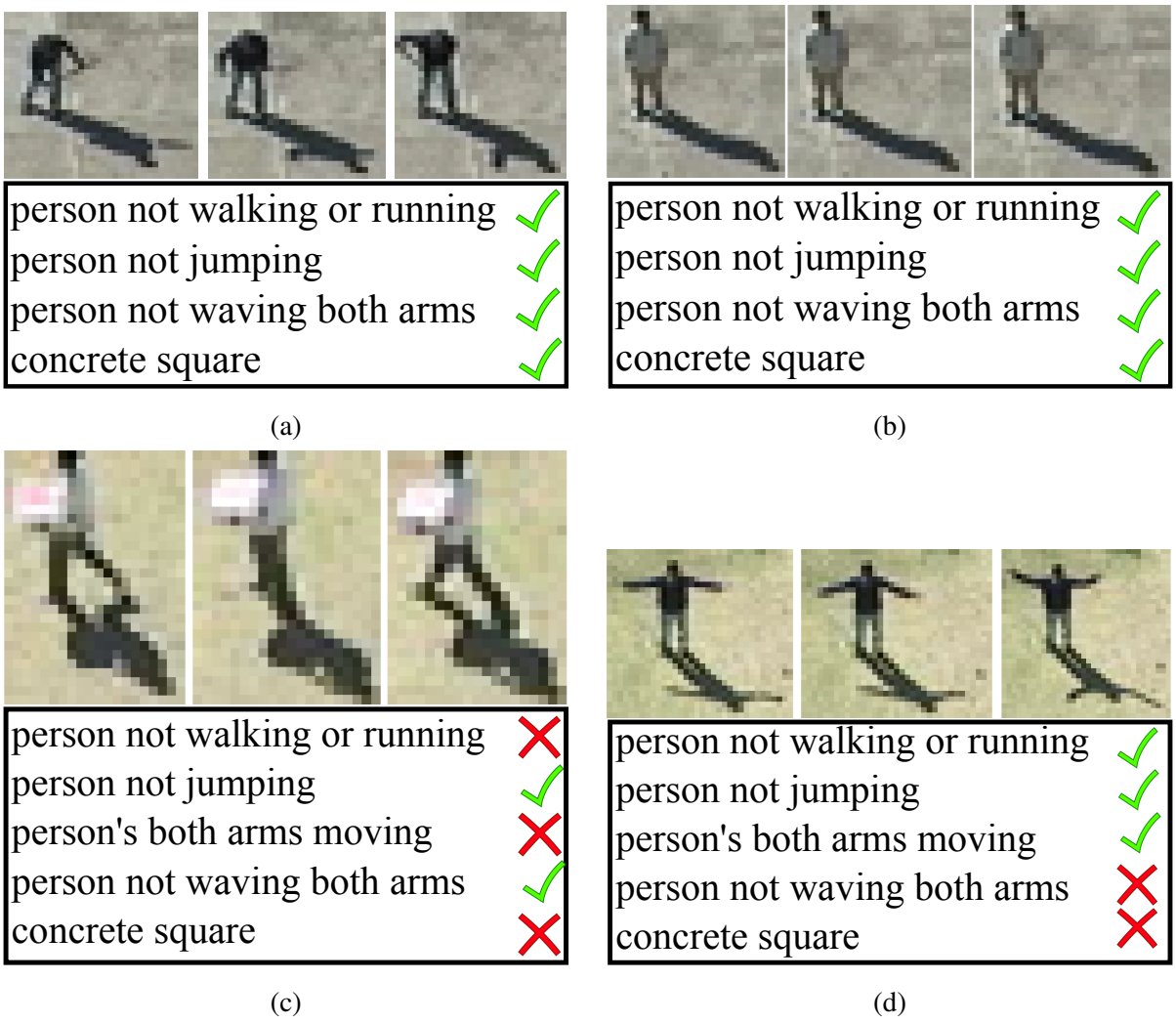
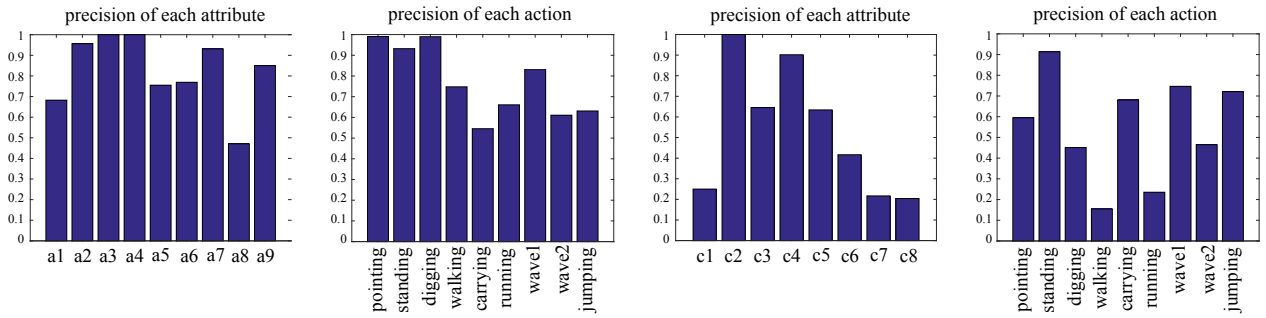


Figure 6.3: The demonstrations of video description based on attributes from PiCoDeS. (a) and (b) are two samples of videos in which most keywords are suitable; (c) and (d) are two of the worst ones.



(a) (b) (c) (d)

Figure 6.4: The detailed results of precision of each attribute and precision of each action for PiCoDeS in (a), (b) and SPH in (c), (d). The horizontal axis in (a), (c) indicates the index of discovered attributes and the axes in (b), (d) indicate the names of actions. The vertical axes represent hit rate (precision); a1 is 'person not walking or running', a2 is 'person not jumping', a3 is 'person lower part stationary', a4 is 'person's four limbs not moving', a5 is 'person's both arms moving', a6 is 'person pointing', a7 is 'person not waving both arms', a8 is 'concrete square', a9 is 'lawn scenes', c1 is 'person standing', c2 is 'person not walking or running', c3 is 'person not moving both arms', c4 is 'person not carrying', c5 is 'person arms not moving separately', c6 is 'person not moving arms', c7 is 'person holds arm in air', c8 is 'person carrying'.

rate with most videos being described with hit rate more than 60%. The results for SPH are worse than PiCoDeS.

6.4.4 Analysis on Cost and Time Saving in the Manual Process

Here we compare the time and cost required to perform manual work between our method and the traditional approaches requiring extensive manual processing. The time and cost analysis is based on the AMT Human Intelligent Task (HIT). One HIT normally comprises a set of tasks that a human could do to label one image/video data. Let J be the number of keywords, and N be the number of training samples which is usually a very large number. In our method, we are only required to name the discovered attributes. Hence, our method requires just J HITs. On the other hand, traditional approaches require at least N HITs as these require all training samples to have the keywords. Indeed as $J << N$, then our method massively reduces the time and cost required as it has much less number of HITs.

6.5 Summary

In this chapter, we described an attribute-based video keyword generation approach. Our approach utilised an existing automatic attribute discovery approach to discover the keywords. Since there

have been numerous attribute discovery approaches in the literature, we devise a selection method, based on the shared structure exhibited amongst meaningful attributes, that enables us to compare the efficacy between different automatic attribute discovery approaches. In particular, we devised a distance function that measures the meaningfulness of a set of discovered attributes. We used our approach to select the methods that are most likely to discover meaningful attributes. Then, we validated our approach on two attribute datasets. The results showed that our approach is able to determine which automatic attribute discovery method can generate the most meaningful keywords or attributes. Finally, we showed how the discovered attributes were used to generate keywords for videos recorded from a surveillance system.

The proposed approach indicates that it is possible to dramatically reduce the amount of manual work in generating video keywords without limiting ourselves to arbitrarily preselected video feature descriptors.

We note that our proposed selection method only indicates the best attribute discovery method. Thus, a more quantitative approach may be required for the future study. In addition, various regularisations such as the ℓ_1 regularisation for (5.1) and (5.2) will be explored in the future. The ℓ_1 constraint is an explicit regularisation to induce sparsity. As to the robustness aspect, our proposed system depends on the robustness of the selected attribute discovery methods. However, further studies on various surveillance datasets are required to fully understand the proposed system robustness.

Chapter 7

Unsupervised Automatic Attribute Discovery Method via Multi-graph Clustering

In a dark time, the eye begins to see.

Cavett Robert

Chapter Summary: *Inspired by the insights from previous work, it is possible to discover attributes from a set of unlabelled data. More importantly, it is also reported in some works that preserving local neighbourhood in the attribute binary space, as well as increasing attribute binary feature discrimination, will improve the meaningfulness. We propose a novel unsupervised attribute discovery method utilising a multi-graph approach that preserves both local neighbourhood structure as well as class separability. Whilst the local neighbourhood structure is preserved by considering multiple similarity graphs, the class separability is achieved by incorporating the traditional clustering objective.*

7.1 Introduction

Inspired by the insights from the design of the meaningfulness metric of the visual attribute, in this chapter, we propose a novel category level unsupervised attribute discovery method utilising multi-graph approach that preserves both local neighbourhood structure as well as cluster separability. The work in this chapter and its related literature reviewed here greatly answer the research question 3 in Section 1.1.

Recently, automatic image and video analytics and description have drawn much attention from the computer vision community [28, 49, 29, 13]. In these research fields, the visual attribute features (high-level human understandable features) have been shown to be powerful tools [28, 49, 51]. For

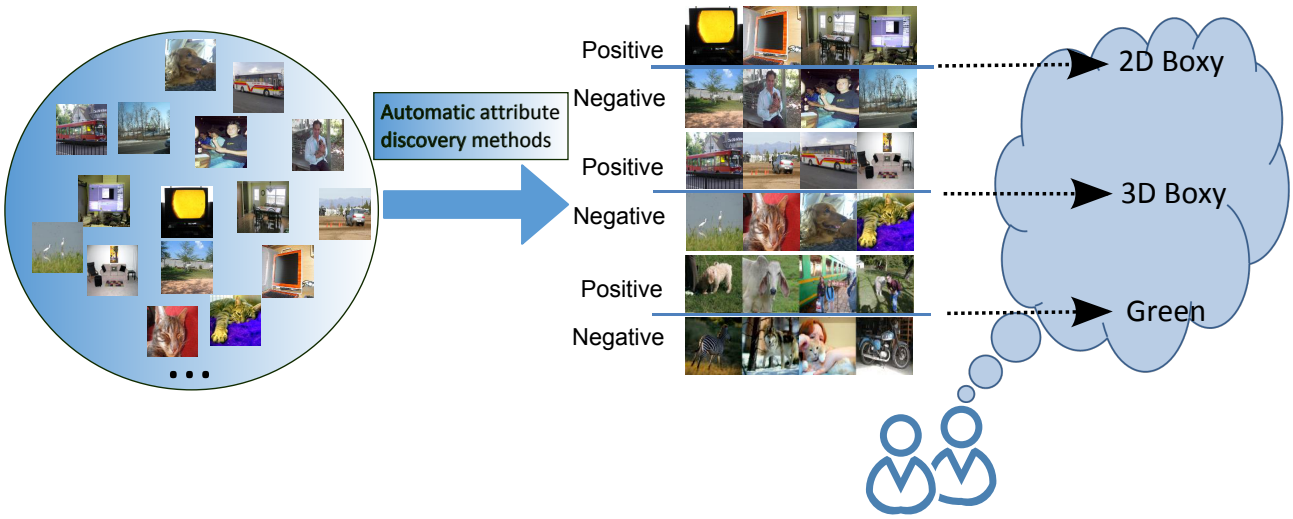


Figure 7.1: Illustration of the concept of automatic attribute discovery. The automatic attribute discovery methods can divide a set of samples into partitions of the meaningful attributes that are potentially human understandable

example, they can be used to perform zero-shot learning where no visual samples of the test categories are given [51].

However, one of the biggest challenges in using attribute descriptors is that a set of labelled images is required to train the attribute classifiers. It is a tedious, time-consuming, and expensive task to label each individual image for every single attribute, especially when there are a large number of images and attributes. Furthermore, the labelling task could be extremely expensive or even impossible in some specialised domains such as *Ornithology* [114], *Entomology* [109], and cell pathology [115] as only highly trained experts are able to complete the work.

To that end, several automatic attribute discovery methods have been developed. These works primarily focus on learning an embedding function that maps the original descriptors into a binary code space wherein each individual bit is expected to represent a visual attribute. These methods are closely related to hashing methods [34, 54, 113]. Nevertheless, unlike hashing methods, which only consider reducing the storage demand and computational complexity whilst maintaining system accuracy, the automatic attribute discovery methods focus on finding meaningful visual concepts in the given set of images. Figure 7.1 illustrates how these methods discover visual attributes.

Since they are supposed to discover meaningful visual attributes, finding the most suitable method that can discover more meaningful attributes is an important research problem. In Chapter 5, we proposed a novel metric to measure attribute meaningfulness. Then we used the proposed metrics to study the efficacy of the attribute discovery methods in finding meaningful attributes. The works provide some insights on how to design automatic attribute discovery methods. For instance, some insights found from these works include: (1) preserving local neighbourhood in the attribute binary space will increase the meaningfulness, and (2) increasing attribute binary feature discrimination will

improve the meaningfulness. In addition, it was also shown that Spectral Hashing (SPH) [113] was able to discover meaningful attributes. This is could be due to the fact the SPH preserves the local neighbourhood in the attribute binary space.

Encouraged by these results, in this chapter, we propose a novel unsupervised automatic attribute discovery method. Our method combines the idea of spectral clustering and spectral hashing methods described in [107, 73]. In particular, the attributes are found through preserving the local neighbourhood by employing multiple similarity graphs and maximising data separability structure by considering a clustering objective function. We use data separability measure instead of feature discrimination, as the feature discrimination measure is only useful for supervised learning scenarios. Furthermore, we consider multiple neighbourhood graphs to capture various local neighbourhood structures. By doing this, our method discovers the attributes by solely studying the underlying geometric structure of the given data. We call our proposed Multi-Graph Clustering Attribute Discovery method, MGCAD. The proposed method can be extremely useful in many real scenarios such as video surveillance data analysis in Chapter 6 and cell pathology [115], where expert knowledge is needed. This means, finding meaningful attributes can be as simple as collecting the data; hence, side-stepping the enormously expensive manual hand labelling process. We evaluate the proposed attribute discovery method using the attribute meaningfulness metric proposed in Chapter 5. The results show the proposed attribute discovery method outperforms recent methods in attribute meaningfulness. Moreover, results from the clustering task evaluation suggest the efficacy of the proposed method compared to Spectral Clustering [113] and the traditional k-means. We list our contribution as follows:

Contributions:

- We propose a novel unsupervised automatic attribute discovery method by considering multiple neighbourhood graphs and clustering objective function.
- Multiple graphs are integrated into this framework to capture different characteristics of the given image exemplars.
- We show that the proposed method outperforms other comparative methods in the clustering task. In addition, when evaluated using the proposed meaningfulness metric in Chapter 5, our approach outperforms recent automatic unsupervised attribute discovery methods.

We continue the chapter as follows. We introduce our proposed attribute discovery method in Section 7.2. Next, we discuss the experiments and results in Section 7.3. Finally, the main findings and future directions are concluded in Section 7.4.

7.2 Approach

In this section, we introduce the details of our proposed MGCAD.

7.2.1 Attribute Representation

Supposed we have a set of samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, a visual attribute can be considered as a decision boundary dividing the set into two subsets $\mathcal{X}^+ \cup \mathcal{X}^- = \mathcal{X}$. Here, $\mathcal{X}^+$ represents the set where the attribute is present and $\mathcal{X}^-$ represents the set where the attribute is absent. Thus, all the attributes are lying on a manifold formed by decision boundaries [77].

As such, an attribute can be represented by an N-dimensional binary vector whose each element is the classification output of sample $\mathbf{x}_i$ tested by the corresponding attribute binary classifier $\phi(\cdot) \in \mathbb{R}$. The sign of the classifier output on $\mathbf{x}_i$ shows whether the sample belongs to the positive or negative set (*i.e.,* $\mathcal{X}^+$ or $\mathcal{X}^-$). In this case, an attribute can be represented as $\mathbf{z}^{[\mathcal{X}]} \in \{-1, +1\}^N$ whose i-th element is $\mathbf{z}_{(i)}^{[\mathcal{X}]} = \text{sign}(\phi(\mathbf{x}_i)) \in \{-1, +1\}$. For simplicity, we drop the symbol $[\mathcal{X}]$ from $\mathbf{z}^{[\mathcal{X}]}$ whenever the context is clear.

7.2.2 The Framework of Single-graph Clustering Attribute Discovery

As suggested in Chapter 5, to increase the attribute meaningfulness, one needs to preserve the local neighbourhood structure and increase the separability of the descriptor. The latter property can be understood as improving the clustering results. One possible way is to rely on graph-based approaches. Spectral hashing [113] provides a practical way of generating good binary code representation of data since it considers using graph a embedding method to preserve the local neighbourhood structure and encode that into the binary code space. However, it has not considered the separability of the data. To overcome this drawback, we propose to optimally jointly learn the objective to preserve the local neighbourhood structure and maximise the separability of the descriptors.

In general, the spectral clustering algorithm can be divided in two main stages: First, one needs to establish a similarity graph, represented in a similarity matrix. Each element in the similarity matrix indicates the connectivity between samples. A new representation that preserves the similarity information encoded in this graph is obtained. This can be done by solving the optimisation problem. Second, the method will perform the clustering task on the new representation space. More specifically, the traditional k-means objective is solved using the k-means algorithm.

Thus, the first stage can be defined via:

$$\max_{F^T F = I} Tr(F^T D^{-\frac{1}{2}} A D^{-\frac{1}{2}} F) \tag{7.1}$$

where $A \in \mathbb{R}^{n\times n}$ is the initial similarity matrix whose elements represent the connections between samples. For example, the element A_{ij} indicates the similarity between sample x_i and x_j. Tr represents the operator of calculating $trace$ of an matrix. $D \in \mathbb{R}^{n\times n}$ represents the degree matrix which is a diagonal matrix with i-th diagonal element as $\sum_j a_{ij}$. $F \in \mathbb{R}^{n\times m}$ represents the projected matrix constructed by the new m-dimensioned vector representation of each sample.

The second stage can be defined as follows:

$$\min_{Y\in Ind,C} \left\|F - YC^T\right\|_F^2 \tag{7.2}$$

where $Y \in$ *Ind* is the indicator matrix that indicates which sample is belonging to which cluster. Here, *Ind* is a set where its element satisfies constraints as $G \in \{0,1\}^{n\times c}$, where each row in G has $||g_{i\cdot}||_0 = 1$; $C \in \mathbb{R}^{m\times c}$ represents the matrix constituted by the vector representation of the centre of each cluster. The number of clusters is c. As a convention in spectral clustering [107], m is usually set as c.

Here, we propose to jointly learn the two stages as (7.3) to get the optimal new representation which preserves the local structure and separability of the data:

$$\max_{F^TF=I,Y\in Ind,C} Tr(F^TD^{-\frac{1}{2}}AD^{-\frac{1}{2}}F) - \eta\left\|F - YC^T\right\|_F^2 \tag{7.3}$$

where η is a parameter.

We use an iterative optimisation method to solve the objective function. First, we apply the first step of spectral clustering method to initialise the objective function, obtaining the projected matrix $F \in \mathbb{R}^{n\times m}$. More specifically, we construct the similarity matrix A and normalise it via $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. We denote this as S. Next, we compute the first m largest eigenvalues of the normalised similarity matrix S. Then corresponding eigenvectors constitute matrix $F \in \mathbb{R}^{n\times m}$. After initialisation, each variable can be computed by an iterative optimisation approach. To that end, if we first fix F, then (7.3) becomes a normal k-means problem as: $\min_{Y\in Ind,C} \left\|F - YC^T\right\|_F^2$. Variables Y and C can be obtained by applying the k-means algorithm.

Then, with the computed Y and C, the optimisation problem of (7.3) can be derived as:

$$\max_{F^TF=I} Tr(F^TD^{-\frac{1}{2}}AD^{-\frac{1}{2}}F) - \eta Tr(F^TYC^T) \tag{7.4}$$

(7.4) can be solved by Generalised Power Iteration [74]. The whole iteration algorithm can be listed in Algorithm 3.

Algorithm 3: The proposed iteration algorithm for solving (7.3).

Input: graph: $A \in \mathbb{R}^{n\times n}$, number of clusters: c, parameters: η, projected dimension: m

1: Initialise $F \in \mathbb{R}^{n\times m}$, which is formed by the m eigenvectors of $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ corresponding to the m largest eigenvalues.
2: **repeat**
3: Update Y, C using k-means algorithm.
4: Update F by solving (7.4) via Generalised Power Iteration [74];
5: **until** converges

7.2.3 The Framework of Multi-Graph Clustering Attribute Discovery

Furthermore, a single graph may not be sufficient to capture the actual relationship between samples. We address this by incorporating different graphs. Here we use three widely used graphs: the Gaussian graph, k-nearest neighbour graph (knn), and the ϵ-neighbourhood graph. In Gaussian graph, we simply connect all samples with each other via a similarity function. Thus, we determine each element in the similarity matrix by using the Gaussian similarity function $A_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. The parameter σ controls the width of the neighbourhoods. In k-nearest neighbour graph, the goal is to connect samples with each other if one sample is among the k nearest neighbours of the other sample. Here we use normal k-nearest neighbour graph, where the connectivity between two samples is established when at least one of the samples is in the k-nearest neighbourhood set of the other sample. We still use Gaussian similarity function with same parameter σ to denote the weight of the edge. In the ϵ-neighbourhood graph, we connect all samples whose pairwise distances are smaller than ϵ. Since the distances between all the samples are at the same scale, we can basically use the unweighted graph which has entry 1 indicating connected between two samples, and entry 0 otherwise. We refer the readers to [107] for full treatment of these graphs.

The final objective function with multi-graphs can be written as (7.5):

$$\begin{aligned}&\max_{\alpha^T\alpha=1,\alpha\geqslant 0,F_v^TF_v=I,Y,C}\sum_{v=1}^{V}\alpha_v Tr(F_v^T D_v^{-\frac{1}{2}}A_v D_v^{-\frac{1}{2}}F_v)\\&-\eta\left\|[F_1,...,F_V]-YC^T\right\|_F^2\end{aligned} \tag{7.5}$$

where α and v represent the mixing coefficient matrix and index of a specific graph respectively. The α_v is the mixing coefficient of the v-th graph.

We can also use an iterative optimisation method to solve the objective function. As in 7.2.2, we initialise the objective function obtaining the projected matrix $F_v \in \mathbb{R}^{n\times m}$. In other words, we construct a similarity matrix A_v for each graph and normalise it via $D_v^{-\frac{1}{2}}A_vD_v^{-\frac{1}{2}}$. We denote this as

S_v. Next, for each graph, we compute the first m largest eigenvalues of the normalised similarity matrix S_v. The projected matrix $F_v \in \mathbb{R}^{n\times m}$ can be constructed by the corresponding eigenvectors.

Then we can start the iterative optimisation approach to compute each variable. To that end, we can compute variables Y and C by solving the k-means problem as: $\min_{Y\in Ind,C} \left\|[F_1, ..., F_V] - YC^T\right\|_F^2$ given F_v and α_v fixed.

Then, we can calculate α_v by plugging the initialised F_v and computed Y and C into (7.5) and get:

$$\max_{\alpha^T\alpha=1,\alpha\geqslant 0} \sum_{v=1}^{V} \alpha_v Tr(F_v^T D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}} F_v). \tag{7.6}$$

where the $\alpha = [\alpha_1, ..., \alpha_V]$.

The reason why we use the orthodox constraint on is because if we only use the $\sum_{v=1}^{V} \alpha_v = 1$, we will encounter a linear programming which can be solved by simplex method. However, in linear programming problem, the optimal solution only appears on the vertices which means only one element in α will be non-zero so that only one graph will be chosen for the clustering task. This apparently contradicts the idea of multi-graph clustering. Therefore the constraint $\alpha^T\alpha = 1$ in (7.6) makes the optimisation problem a quadratic programming problem which is a better approximation of the real case.

To solve (7.6) and obtain α_v, we need the following lemma.

Lemma 7.2.1. *Given any vector $\beta = [\beta_1, ..., \beta_V]$, the solution of the following optimisation problem: $\max_{\alpha^T\alpha=1,\alpha\geqslant 0} \sum_v \alpha_v\beta_v$ is given by $\alpha_v = \frac{\beta_v}{\sqrt{\sum_v \beta_v^2}}$*

Proof. This problem can be proved by the Lagrange method. We can introduce a new variable λ, the optimisation problem can then be derived as a Lagrange function $\mathcal{L}(\alpha, \lambda) = \alpha^T\beta - \lambda(\alpha^T\alpha - 1)$. Taking derivation on both sides, we will have $\mathcal{L}(\alpha, \lambda)' = \beta - \lambda\alpha$. Setting $\mathcal{L}(\alpha, \lambda)' = 0$, then we will have $\alpha_v = \frac{\beta_v}{\lambda}$. Considering the constraint $\alpha^T\alpha = 1$, then the solution is $\lambda = \sqrt{\sum_k \beta_k^2}, \alpha_v = \frac{\beta_v}{\sqrt{\sum_k \beta_k^2}}$ □

According to Lemma 7.2.1, we can obtain the solution as:

$$\alpha_v = \frac{Tr(F_v^T D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}} F_v)}{\sqrt{\sum_v Tr(F_v^T D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}} F_v)^2}} \tag{7.7}$$

Algorithm 4: The proposed iteration algorithm for solving (7.5).

Input: graphs: $A = [A_1, ...A_V]$ whose arbitrary component $A_v \in \mathbb{R}^{n\times n}$, number of clusters: c, parameters: η, projected dimension m

1: Initialise $F_v \in \mathbb{R}^{n\times m}$, which is formed by the m eigenvectors of $D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}}$ corresponding to the m largest eigenvalues.
2: **repeat**
3: Update Y, C using k-means algorithm.
4: Update coefficient α_v according to (7.7)
5: Update F by solving (7.8) via Generalised Power Iteration [74];
6: **until** converges

Finally to optimise the F_v with fixed Y, C, and α, the objective function (7.5) can be derived as follows:

$$\begin{aligned} &\max_{F_v^T F_v = I} \sum_{v=1}^{V} \alpha_v Tr(F_v^T D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}} F_v) - \eta \left\| [F_1, ..., F_V] - YC^T \right\|_F^2 \\ \Rightarrow &\max_{F_v^T F_v = I} \sum_{v=1}^{V} \alpha_v Tr(F_v^T D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}} F_v) + 2\eta Tr([F_1, ..., F_V]^T YC^T) \\ \Rightarrow &\max_{F_v^T F_v = I} \alpha_v Tr(F_v^T D_v^{-\frac{1}{2}} A_v D_v^{-\frac{1}{2}} F_v) + 2\eta Tr(F_v^T YC^T) \end{aligned} \tag{7.8}$$

The derived form of the problem in (7.8) can still be solved by Generalised Power Iteration [74].

The specific steps of the algorithm are presented in Algorithm 4.

The computational complexity of the algorithm includes three main parts. The first part lies on the step 3 in Algorithm 4, which uses the k-means algorithm to compute the current Y and C. The computational complexity is $O(n^{dk+1})$ where n is the total number of samples, d is the total number of dimensions which is the sum of all the dimensions from F_v. The second part lies on the step 4 in Algorithm 4, two matrix multiplication operations are performed here. The computational complexity is $O(\sum_{v=1}^{V} \left(d\left(v\right) n^2 + d(v)^2 n\right))$ where V is the number of graphs $d(v)$ is the number of dimension of F_v. The third part lies on the Generalized Power Iteration algorithm which solves the problem in (7.8). The computational complexity is $O(\sum_{v=1}^{V} \left(d(v)^2 n + d(v)^2 qt\right))$ where q is the dimension of new representation of F_v, t is the number of iteration. Therefore, the total computational complexity is $O(\tau \left(n^{dk+1} + \sum_{v=1}^{V} \left(d\left(v\right) n^2 + d(v)^2 n\right) + \sum_{v=1}^{V} \left(d(v)^2 n + d(v)^2 qt\right)\right))$ where τ is the number of iteration of the whole algorithm. The running time of the proposed algorithm on MSRCV1 for 100 iterations is 13.7 seconds. However, the algorithm converges very fast (within 10 iteration normally).

7.2.4 Attribute Generation

After we extract the new representation of data, we still need to encode this into a binary form. Since spectral hashing provides an effective way for doing this, we opt to employ this to perform the task. Here, the spectral hashing is considered as the post-processing stage to obtain the binary code from the new representation F_v. More specifically, we concatenate the new data representation F_v from each graph and perform PCA to obtain the eigenvectors $U \in \mathbb{R}^{n \times c}$, then we follow the method of spectral hashing to obtain the binary code representation as $\mathbf{z}^{[\mathcal{X}]} \in \{-1, +1\}^N$.

7.3 Experiment

In this section, we first validate the performance of the proposed algorithm in clustering. Then we evaluate the efficacy of the proposed method in discovering attributes. In the clustering task, we use k-means and spectral clustering [107] methods as the baseline comparisons. We then apply these methods on a recent benchmarking clustering dataset, MSCRV1 [117, 108, 9]. In attribute discovery evaluation, we contrast our proposed method using a well-known public dataset a-Pascal a-Yahoo dataset (ApAy) [28] with a recent automatic attribute discovery methods such as PiCoDeS [4] as well as some recent hashing methods. These methods are Iterative Quantisation (ITQ) [34], Spectral Hashing (SPH) [113], and Locality Sensitivity Hashing (LSH) [54].

7.3.1 Experiment Setting

MSRC-v1 [1] **—** comprises 8 classes, 240 images in total. Following [53], we select 7 classes composed of tree, building, airplane, cow, face, car, and bicycle with 30 samples in each class. We extract 6 visual features from each image. They are respectively Color moment, LBP, HOG, SIFT, GIST, Gentrist.

a-Pascal a-Yahoo dataset (ApAy) [28] — comprises two sources: a-Pascal and a-Yahoo. There are 12,695 cropped images in a-Pascal that are divided into 6,340 for training and 6,355 for testing, with 20 categories. The a-Yahoo set has 12 categories disjoint from the a-Pascal categories. Moreover, it only has 2,644 test exemplars. There are 64 attributes provided for each cropped image. The dataset provides four features for each exemplar: local texture, HOG, edge, and colour descriptor. We use the training set for discovering attributes and we perform our study on the test set. More precisely, we consider the test set as the set of images $\mathcal{X}$ defined in Section 7.2.

For the attribute discovery experiment, we apply the following pre-processing step described in [4]. We first lift each feature into a higher-dimensional space which is three times larger than the original space. After the features are lifted, we then apply PCA to reduce the dimensionality

[1] http://research.microsoft.com/en-us/projects/objectclassrecognition/

of the feature space by 40 percent. This pre-processing step is crucial for PiCoDeS as it uses lifted feature space to simplify their training scheme while maintaining the information preserved in the Reproducing Kernel Hilbert Space (RKHS). Therefore, the method's performance will be severely affected when lifted features are not used.

Each method is trained with the training images to discover the attributes. Then we use the manifold $\mathcal{M}$ w.r.t. the test images for the evaluation. More precisely, each attribute descriptor is extracted from test images (*i.e.,* $\mathbf{z}_k, \mathbf{z}_k \in \{-1, 1\}^N$, where N is the number of test images).

7.3.2 Clustering Evaluation

We first perform the clustering experiment to evaluate the performance of the proposed method in a clustering task. We randomly choose the parameters σ, ϵ, and η in a range of $[10^{-4}, 10^{-3}, ..., 10^{4}, 10^{4}]$ and we set the parameter k for the knn graph to 10. The method will perform the optimisation iteration until it converges. We evaluate all the comparative methods on 7 clusters. The results measured by clustering accuracy and normalised mutual information [18] are reported in Table 7.1.

Table 7.1: Clustering performance comparison measured by clustering accuracy (ACC) and normalised mutual information (NMI).

	k-means	Spectral Clustering[113]	MGCAD (proposed)
ACC	0.7462	0.7789	**0.7952**
NMI	0.3561	0.3975	**0.4208**

From the results, we can see the proposed MGCAD outperforms all of the baselines. This indicates multiple different graphs are able to capture different data local neighbourhood structure and reflect the complex relation between samples; thus, increasing the clustering performance.

7.3.3 Attribute Meaningfulness Evaluation

We evaluate our proposed MGCAD method for the attribute discovery task. All of the parameters are set as in Section 7.3.2. Because of the reconstruction framework of the metric, (we need to put aside a part of groundtruth attribute labels as the meaningful subspace to calibrate the score) the number of discovered attribute is better to set less than the number of the groundtruth attributes which are used to construct the meaningful subspace. Empirically, the size of 16, 32 and 64 attribute output are the typical settings in the experiment. However, there are only 64 groundtruth attributes in test dataset. We should choose more than 32 groundtruth attributes for the calibration. Thus it is appropriate to configure each method to discover 16 attributes in the experiments.

For this task we use the attribute meaningfulness metric proposed in Chapter 5. The attribute meaningfulness metric is an automatic quantitative evaluation method which can be considered as a

yardstick for evaluating meaningfulness of a set of automatically discovered attributes. The evaluation method is based on reconstruction error with two different regularisations approximating the geodesic distance between a given attribute set and the meaningful subspace of human labelled attributes. The final metric score is based on subspace interpolation on a decision boundary manifold. In our evaluation, we employ both γ_{cvx} and γ_{jp}, then we average them to get the final metric score $\tilde{\gamma}$ as suggested in Chapter 5.

Here, we first show the computed coefficient α to demonstrate the different graphs will have different contributions to the final results with respect to the weight coefficient α. Since there are four graphs in use for the clustering task, the coefficient vector α has 4 components. The coefficient vector α is shown in Table 7.2.

Table 7.2: The values of coefficient α

	Gaussian	knn	ϵ-neighborhood	CLR
MSRCV1	0.505	0.469	0.512	0.512
ApAy	0.654	0.092	0.657	0.364

The results demonstrate that the coefficient of each graph has been alternatively optimised to have different value and shows different contribution to the final result from each individual graph. Moreover, the differences between different α_v are not too large not too small and structure in α is not sparse. These indicate that each graph can contribute the final results but the contributions are not equal. These observations all indicates that each graph is important to the final result and the multi-graph scheme works.

The final attribute meaningfulness scores $\tilde{\gamma}$ are shown in Table 7.3.

Table 7.3: Attribute meaningfulness comparisons between various attribute discovering methods and hashing methods.

	Scheme used	$\tilde{\gamma}$
LSH [54]	Unsupervised	4.5
ITQ [34]	Unsupervised	10.4
SPH [113]	Unsupervised	16.8
MGCAD (proposed)	Unsupervised	**79.3**
PiCoDeS [4]	Supervised	97.7

From the results, the attribute set discovered by LSH has the lowest meaningfulness score. This is because LSH employs the random projection scheme to generate every single bit of the code. PiCoDeS has a higher attribute meaningfulness score indicating it is more likely to discover more meaningful attributes than other methods. PiCoDeS applies category labels and max-margin framework to jointly learn the category classifier and attribute descriptor in an attempt to maximise the

discriminative power of the descriptor. Note that PiCoDeS is not an unsupervised method since it also utilises the class labels. Among all the unsupervised methods *i.e.,* LSH, ITQ, SPH, MGCAD, our proposed MGCAD outperforms the others. SPH is also a graph based method that is able to discover meaningful attribute to some extent. We conjecture that the performance increase gained by MGCAD compared to SPH is due to two following aspects: (1) it employs multiple graphs; and (2) it considers the data separability via the clustering objective.

7.4 Summary

In this chapter, we proposed a novel unsupervised attribute discovery method by combining multiple similarity graphs and clustering objective. We considered integrating multiple different graphs into this framework to capture different characteristics of given image exemplars. The experiment of clustering task indicates the proposed new framework can help to obtain better data representation, reflecting the separability and neighbourhood preservation. The experiment results of attribute discovery task showed that the proposed method outperforms recent attribute discovery methods. The results further corroborated previous findings in designing effective attribute discover methods in Chapter 5.

In future work, we will further investigate more advanced techniques of graph construction.

Chapter 8

Conclusions and Future Work

I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right.

Albert Einstein

Chapter Summary: *Discriminative power and semantic meaningfulness are two significant properties of the visual attribute. Until now, there is no research thoroughly analysing and carefully considering both of them in practical terms. This research is one of the very few works that starts to consider them in the same framework, and utilise them together to attack some real-world problems. However, there is still a lot of effort required to automatically discover more discriminative, yet meaningful, attributes and explore their novel applications.*

8.1 Thesis Summary

Visual attributes are considered as the high level semantic representation describing inherent property characteristics of images or visual information.

Visual attributes are extremely useful as they are: (1) human understandable; (2) machine computable, and (3) shared across classes.

Despite their advantages, visual attributes still suffer from a number of shortcomings: (1) the need for extensive labelling to train the attribute detectors, and (2) the lack of discriminative power. First, traditional supervised attribute detectors/classifiers [50, 80] are actually concept classifiers, so they need extensive human labelling effort.

In this thesis, we found that the crux of these problems is that the previous works have not yet fully considered the discriminative power and meaningful property of attributes. Therefore in this

research, we put the visual attribute research into the framework of these two properties, focussing on the automatic measuring of the meaningfulness of automatically discovered attributes. Furthermore, we investigate its practical application.

The results of the experiments and the theoretical analysis have shown that the visual attribute is a promising tool in many visual applications such as zero-shot learning in Chapter 4, meaningful binary code generation in Chapter 5, as well as automatic analysis and keyword generation for surveillance video in Chapter 6.

Let us revisit the goal and research questions to conclude the main findings of this thesis. Due to the issues with current visual attribute research, the first research question posed is:

Q1: Can we automatically discover discriminative visual attributes that are still meaningful from images and video data?

Chapters 4-6, and 7 answers this question from different aspects. Chapter 4 answers this question from the pre-defined human labelled attributes. We propose a novel greedy algorithm called Discriminative and Reliable Attribute Learning which selects a subset of pre-defined attributes which maximises an objective function incorporating the two properties. The work in Chapter 5 reviews various recent automatic attribute discovering methods and investigates the semantic meaningfulness property of the visual attribute. From the comparison, we can conclude that some methods can automatically discover more meaningful visual attributes. This, to some extent, shows that it is possible to find a way to automatically discover discriminative and semantically meaningful attributes from image and video data. After that, in this thesis we devise an attribute meaningfulness metric based on the shared structure assumption [77] which will answer research question 2. The work in Chapter 6 performs an application study on the attribute meaningfulness evaluation method proposed in Chapter 5 and further indicates that, not only can the discriminative and semantically meaningful attributes be extracted from videos, they can also be used for keyword generation application in the surveillance video scenario. From the insights of Chapter 5, we devise a novel clustering based unsupervised attribute discovery method which can discover discriminative, yet semantically meaningful, attributes, further answering this question. In practical use, this kind of attribute can serve as better human and computer interaction interface so as to help visual information retrieval such as online shopping and accurate user recommendation, especially in fine grain case. For example, the discovery of local fine-grained attributes that are both discriminative and potentially semantically meaningful, in an unsupervised fashion, further eases the human interactive labelling effort [26].

Q2: Can we automatically discover visual attributes that are meaningful from a set of videos/images with category labels? If yes, how can the attribute meaningfulness be measured?

The literature reviewed in Chapter 2 provided evidence that many discriminative code generation methods such as DBC [86] and PiCoDeS [4] have the ability to discover meaningful attributes. The experiment results in Chapter 5 show that the methods can benefit from the usage of category labels for meaningful attribute discovery. As to the sub-question "how can the attribute meaningfulness be measured?", the work in Chapter 5 thoroughly answers this question through a semantic meaningfulness metric of the visual attribute. With this metric, automatic quantitative evaluation can be performed on the attribute sets; thus, reducing the enormous effort to perform the manual evaluation. The proposed metric is applied to some recent automatic attribute discovery and hashing methods on four attribute datasets. Extensive experiments and results show that the final metric score indicates how much meaningful content is contained within the set of discovered attributes. This answers the question and it is possible to measure the attribute semantic content without much human effort required. In practice, this metric is promising to become a useful tool to largely save human effort and time cost as a pre-procedure for selecting potentially meaningful attributes discovered by other methods before passing to humans for naming them. Therefore it can assist the automation of automatic attribute discovery.

Q3: Can we automatically discover visual attributes from a set of videos/images without any supervision?

The methods reviewed in Chapter 5, such as spectral hashing [113], and Chapter 7 both answer this question. Through the in-depth research on the meaningfulness metric of the automatically discovered attribute, we have shown the evidence that it is possible to measure attribute semantic content without human involvement. Another important finding is that through preserving local neighbourhood structure, it is possible to discover attributes without supervision. However, the discriminative power between classes could also be an important part of discovery of meaningful attributes. Therefore, we consider using the clustering based methods to introduce the separability of clusters, inspired by the discriminative power. The results show that the novel method can discover more meaningful attributes in a totally unsupervised fashion. In this real-world application, attribute-based dataset construction will benefit from this research. For example, this proposed method can assist annotation of raw collected image data without any labels, and mine the potentially meaningful attribute information from the raw data before the real human labelling procedure, so as to speed up and facilitate the visual dataset construction.

8.2 Contribution

This section echoes Section 1.2. The work in this thesis first starts from the discriminative power of visual attributes. Not all visual attributes are discriminative enough to distinguish between classes.

For example, the attribute "fluffy" is not able to distinguish between dogs and cats. Therefore a Discriminative and Reliable Attribute Learning (DRAL) method is proposed to select discriminative, yet meaningfully, reliable attributes from a set of meaningful attribute.

The discriminative power is not the only component that contributes to the attribute application. Meaningfulness of an attribute also plays an important role. For a traditional supervised attribute discovery method, the meaningful property is reflected by attribute prediction reliability, in other words, the generalisation error of the predefined attribute detectors. To that end, in the supervised case with the pre-defined attribute labels, we propose a Discriminative and Reliable Attribute Learning method to select the visual attributes that are most discriminative and reliable so as to achieve better performance in classification tasks.

Recently, many new automatic attribute discovery methods [34, 54, 113] are proposed. The primary aim of these works is to learn a function that maps the original image feature space into a binary code space wherein each individual bit represents the presence/absence of a visual attribute. In contrast to the human predefined attributes, the meaningfulness property is essential in this case because when the mapping function is designed just to increase the discriminative power in the output feature space, the individual binary bits (*i.e.,* one binary bit for one attribute) may not have any relationship with any semantic meaning; hence not meaningful. In order to study this further, extensive human effort is required to evaluate the meaningfulness of each binary bit or each attribute. Due to the tedious and time-consuming nature of traditional human involved evaluation methods, we propose a novel automatic meaningfulness metric that will become one of the *yardsticks* to measure attribute meaningfulness based on *shared structure* assumption [76, 77].

With the help of meaningfulness metric, in this thesis, we also select the method that can automatically discover the most meaningful attributes. We apply this method to study the keyword generation problems in surveillance video scenarios and obtain better results over other attribute discovery methods.

Finally, inspired by the insights from designing the metric, we propose a novel category-level unsupervised attribute discovery method utilising a multi-graph approach that preserves both local neighbourhood structure, as well as class separability.

These four works are:

1. Discriminative and Reliable Attribute Learning method on visual attribute selection for better attribute-based classification performance and efficiency. The proposed method significantly reduces the number of the attributes (only 55 discriminative and reliable attributes used) to achieve similar performance compared to the 85 attributes used in DAP. We can also reduce the number of attributes used to 40 with a price of slight performance loss (from 41.5% to 41.2%). The results suggest that we are able to discover the discriminative attributes from the semantic attribute set.

2. A novel automatic meaningfulness metric is proposed to measure attribute meaningfulness based on *shared structure*. To that end, we proposed a novel metric, here called the attribute meaningfulness metric. We developed two distance functions for measuring the meaningfulness of a set of attributes. The distances were then calibrated by using subspace interpolation between Meaningful Subspace and Non-meaningful/Noise Subspace. In the extensive experiment, our metric was also shown to be more correlated with the user study compared with a metric adapted from a recent semi-supervised attribute discovery method, especially the experiment of the logarithmic fitting on the relation between the score proposed metric and the user study compared with that of the MPPCA method (0.99 vs 0.64 on ApAy dataset, and 0.98 vs 0.89 on ASUN dataset, using the coefficients of determination R^2). All results suggested that there is a strong indication that the shared structure may exist among the meaningful attributes. The results also suggest that discovering attributes by optimising the attribute descriptor discrimination and/or preserving the local similarity structure could yield more meaningful attributes.

3. A keyword generation application in surveillance video scenarios is proposed based on the automatic attribute discovery method chosen by the attribute meaningfulness measurement. The results showed that our approach is able to determine which automatic attribute discovery method can generate the most meaningful keywords or attributes. Then, we showed how the discovered attributes were used to generate keywords for videos recorded from a surveillance system. The proposed approach indicates that it is possible to dramatically reduce the amount of manual work in generating video keywords without limiting ourselves to arbitrarily preselected video feature descriptors.

4. A novel category-level unsupervised attribute discovery method via clustering framework combining multiple similarity graphs is proposed to automatically discover visual attributes. We considered integrating multiple different graphs into this framework to capture different characteristics of given image exemplars. The clustering indicates the proposed new framework can help to obtain better data representation reflecting the separability and neighbourhood preservation (at least $\tilde{2}\%$ better in ACC, and $\tilde{2}.3\%$ better in NMI, than other comparative methods). The experiment results of attribute discovery task showed that the proposed method outperforms recent attribute discovery methods (~50 better in quantitative value using attribute meaningfulness than other unsupervised attribute discovery methods). The results further corroborated previous findings in designing effective attribute discover methods in Chapter 5 and 6.

8.3 Future Work

There are many extensions and feasible enhancements that can be explored in the future. For instance, in Chapter 4, we could use a better approximation to measure the attribute reliability property that considers both the detector performance as well as the semantic drift. Another interesting future direction is to find the smallest set of attributes by adding an additional regularisation term in the objective function. We can also explore some novel applications for the proposed strategy such as super resolution [59], 3D reconstruction [127, 128], or anomaly detection in surveillance systems [116]. Here, we can use attributes of low-resolution images as the query to collect the high-resolution images which have parts similar to that, then use the patches of the high-resolution images as sources to approximate the patches of low-resolution image and reconstruct the high-resolution images.

For Chapter 5, we can explore other constraints or optimisation models to capture the hierarchical structure of semantic concepts. A more large-scale user study using AMT on other datasets may be more desirable since human evaluation also plays an important part in further validation. Some other directions could be to investigate the influence of the degenerated or low-resolution image [59] on the attribute meaningfulness evaluation, or to evaluate the potential attributes for 3D reconstructed image sequences [127]. Furthermore, The same meaningful subspace approximation idea can be extended using a deep learning framework. As the features generated in deep learning layers such as fully CNN are purportedly containing low-level semantics, the meaningful subspace approximation can also be performed using these features. The approximation may be better via this way, since the number of features with independent semantics generated from CNN could be large.

For the application study in Chapter 6, we note that our proposed selection method only indicates the best attribute discovery method. Thus, a more quantitative approach may be required for the future study. In addition, various regularisations such as the $\ell_{2,1}$ regularisation for (5.1) and (5.2) will be explored in the future. The $\ell_{2,1}$ constraint is an explicit regularisation to induce sparsity. As to the robustness aspect, our proposed system depends on the robustness of the selected attribute discovery methods; however, further studies on various surveillance datasets are required to fully understand the proposed system robustness.

Moreover, in Chapter 7, we can further investigate more advanced techniques of graph construction.

In more profound discussion, the thesis might be related to the bouba/kiki effect in the sense of how humans will name an object. The bouba/kiki effect [1] is a non-arbitrary mapping between speech sounds and the visual shape of objects. This effect was first observed in psychology research [84]. Basically, the effect suggests that the human brain somehow attaches abstract meanings to the shapes and sounds in a consistent way. For example, the rounded shape may most commonly be named ”bouba” because the mouth makes a more rounded shape to produce that sound while a more taut, angular mouth shape is needed to make the sound ”kiki”. Thus, the bouba/kiki effect has implica-

tions for the evolution of language, because it suggests that the naming of objects is not completely arbitrary. From this clue, we can even incorporate the future unsupervised visual attribute discovery method with the prior knowledge of the effect as a constraint to better approximate the human perception so as to discover attributes with more meaningfulness.

Bibliography

[1] Bouba/kiki effect. `http://en.wikipedia.org/wiki/Bouba/kiki_effect`.

[2] H. S. Baird, A. L. Coates, and R. J. Fateman. Pessimalprint: a reverse turing test. *International Journal on Document Analysis and Recognition*, 5(2-3):158–163, 2003.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[4] A. Bergamo, L. Torresani, and A. W. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*, 2011.

[5] R. Bishop and S. Goldberg. *Tensor Analysis on Manifolds*. Dover Publications, 2012.

[6] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013.

[7] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CVIR*, 2007.

[8] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[9] X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. In *IJCAI*, 2013.

[10] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[11] X. Chang, F. Nie, Y. Yang, and H. Huang. A convex formulation for semi-supervised multi-label feature selection. In *Proc. AAAI*, 2014.

[12] X. Chang, H. Shen, S. Wang, J. Liu, and X. Li. Semi-supervised feature analysis for multimedia annotation by mining label correlation. In *PAKDD*, 2014.

[13] X. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016.

[14] O. Chapelle. Training a support vector machine in the primal. *Neural computation*, 19(5):1155–1178, 2007.

[15] S. S. Chawathe and H. Garcia-Molina. Meaningful change detection in structured data. In *ACM SIGMOD Record*, volume 26, pages 26–37. ACM, 1997.

[16] C.-C. Chen, M. S. Ryoo, and J. K. Aggarwal. Ut-tower dataset: Aerial View Activity Classification Challenge, 2010. `http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html`.

[17] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision (ECCV)*, pages 609–623. 2012.

[18] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *Twenty-Fifth Conference on Artificial Intelligence*, 2011.

[19] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *(ICCV)*, 2013.

[20] J. Chung, D. Lee, Y. Seo, and C. D. Yoo. Deep attribute networks. In *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, volume 3, 2012.

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *(CVPR)*, 2005.

[22] M. Das, S. Amer-Yahia, G. Das, and C. Yu. Mri: Meaningful interpretations of collaborative ratings. *Proceedings of the VLDB Endowment*, 4(11), 2011.

[23] J. Deng, J. Krause, M. Stark, and L. Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):666–676, 2016.

[24] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1657–1664, 2011.

[25] N. R. Draper and H. Smith. *Applied regression analysis*. John Wiley & Sons, 2014.

[26] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3474–3481, 2012.

[27] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2352–2359, 2010.

[28] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[29] J. Feng, S. Jegelka, S. Yan, and T. Darrell. Learning scalable discriminative dictionary with sample relatedness. In *CVPR*, 2014.

[30] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. 2014.

[31] V. Ferrari and A. Zisserman. Learning visual attributes. 2008.

[32] Y. Fu, S. Gong, T. M. Hospedales, and T. Xiang. Learning multi-modal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2013.

[33] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*. 2012.

[34] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[35] R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11):2288–2302, 2014.

[36] S. Hong, J. Choi, J. Feyereisl, B. Han, and L. S. Davis. Joint image clustering and labeling by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1411–1424, 2016.

[37] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on*, 16(4):1168–1181, 2007.

[38] C. Huang, C. Change Loy, and X. Tang. Unsupervised learning of discriminative attributes and visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[39] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1761–1768, 2011.

[40] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, pages 2429–2437, 2011.

[41] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011.

[42] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*. 2012.

[43] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *(CVPR)*, 2012.

[44] T. Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, 2008.

[45] A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision*, pages 1–18, 2015.

[46] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210, 2015.

[47] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1403–1410, 2011.

[48] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[49] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

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

[51] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:1, 2013.

[52] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, volume 2, page 3, 2012.

[53] Y. J. Lee and K. Grauman. Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, 85(2):143–166, 2009.

[54] J. Leskovec, A. Rajaraman, and J. Ullman. *Mining of Massive Datasets*. Cambridge university press, 2013.

[55] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[56] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos. Recognizing activities via bag of words for attribute dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2587–2594, 2013.

[57] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344, 2011.

[58] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In *WACV*, pages 339–346, 2013.

[59] L. Liu, W. Li, S. Tang, and W. Gong. A novel separating strategy for face hallucination. In *IEEE International Conference on Image Processing (ICIP)*, 2012.

[60] L. Liu, F. Nie, T. Zhang, A. Wiliem, and B. C. Lovell. Unsupervised automatic attribute discovery method via multi-graph clustering. In *ICPR*, 2016.

[61] L. Liu, A. Wiliem, S. Chen, and B. C. Lovell. Automatic image attribute selection for zero-shot learning of object categories. In *ICPR*, 2014.

[62] L. Liu, A. Wiliem, S. Chen, and B. C. Lovell. Automatic and quantitative evaluation of attribute discovery methods. In *WACV*, 2016.

[63] L. Liu, A. Wiliem, S. Chen, and B. C. Lovell. Determining the best attributes for surveillance video keywords generation. In *The IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2016.

[64] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *CVPR*, 2012.

[65] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, 40(1):262–282, 2007.

[66] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[67] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[68] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[69] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann. Complex event detection via multi-source video attributes. In *(CVPR)*, 2013.

[70] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010.

[71] B. Manthey and R. Reischuk. The intractability of computing the hamming distance. *Theoretical Computer Science*, 337(13):331 – 346, 2005.

[72] A. H. Meghdadi and P. Irani. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2119–2128, 2013.

[73] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 2001.

[74] F. Nie, R. Zhang, and X. Li. A generalized power iteration method for solving quadratic problem on stiefel manifold. In *arXiv*, 2016.

[75] 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.

[76] D. Parikh and K. Grauman. Interactive discovery of task-specific nameable attributes. In *Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.

[77] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[78] D. Parikh and K. Grauman. Relative attributes. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

[79] A. Parkash and D. Parikh. Attributes for classifier feedback. In *(ECCV)*. 2012.

[80] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[81] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[82] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

[83] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg. Clustered synopsis of surveillance video. In *AVSS*, pages 195–200, 2009.

[84] V. S. Ramachandran and E. M. Hubbard. Synaesthesia–a window into perception, thought and language. *Journal of consciousness studies*, 8(12):3–34, 2001.

[85] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. In *ICCV*, 2013.

[86] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. 2012.

[87] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. pages 184–195, 2014.

[88] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition (GCPR 2014)*. Springer, 2014.

[89] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *European Conference on Computer Vision (ECCV)*, pages 144–157. 2012.

[90] M. Rohrbach, Q. Wei, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013.

[91] Y. Rui and Z. Liu. Artifacial: Automated reverse turing test using facial features. *Multimedia Systems*, 9(6):493–502, 2004.

[92] P. Samangouei, V. M. Patel, and R. Chellappa. Attribute-based continuous user authentication on mobile devices. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8, 2015.

[93] A. Sanin, C. Sanderson, M. Harandi, and B. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *WACV*, 2013.

[94] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426, 2007.

[95] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*. 2012.

[96] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[97] C. Silberer, V. Ferrari, and M. Lapata. Models of semantic representation with visual attributes. In *ACL*, 2013.

[98] Y. Su and F. Jurie. Learning compact visual attributes for large-scale image classification. In *ECCV Workshops and Demonstrations*, 2012.

[99] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[100] A. M. Turing. Computing machinery and intelligence. *Mind*, pages 433–460, 1950.

[101] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008.

[102] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[103] V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000.

[104] J. Varadarajan, R. Emonet, and J.-M. Odobez. A sequential topic model for mining recurrent activities from long term video logs. *International journal of computer vision*, 103(1):100–126, 2013.

[105] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, March 2012.

[106] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *Advances in CryptologyEUROCRYPT 2003*, pages 294–311. 2003.

[107] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[108] H. Wang, F. Nie, and H. Huang. Multi-view clustering and feature learning via structured sparsity. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

[109] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.

[110] W. Wang, Y. Yan, S. Winkler, and N. Sebe. Category specific dictionary learning for attribute specific feature selection. *IEEE Transactions on Image Processing*, 25(3):1465–1478, 2016.

[111] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539–555, 2009.

[112] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Computer Vision–ECCV 2010*, pages 155–168. 2010.

[113] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2009.

[114] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[115] A. Wiliem, P. Hobson, and B. C. Lovell. Discovering discriminative cell attributes for hep-2 specimen image classification. In *WACV*, 2014.

[116] A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda. A suspicious behaviour detection using a context space model for smart surveillance systems. *Computer Vision and Image Understanding*, 116(2):194–209, 2012.

[117] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005.

[118] F. Xiao and Y. Jae Lee. Discovering the spatial extent of relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[119] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[120] X. Xu, T. Hospedales, and S. Gong. Discovery of shared semantic spaces for multi-scene video query and summarization. *arXiv preprint arXiv:1507.07458*, 2015.

[121] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. Action recognition by learning bases of action attributes and parts. In *(ICCV)*, 2011.

[122] A. Yu and K. Grauman. Just noticeable differences in visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[123] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[124] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*. 2010.

[125] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[126] K. Zhao, A. Wiliem, and B. Lovell. Kernelised orthonormal random projection on grassmann manifolds with applications to action and gait-based gender recognition. In *Identity, Security and Behavior Analysis (ISBA), 2015 IEEE International Conference on*, 2015.

[127] Y. Zhu, D. Huang, F. De La Torre, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *(CVPR)*, 2014.

[128] Y. Zhu and S. Lucey. Convolutional sparse coding for trajectory reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2013.