Semantic Structure Discovery in Surveillance Videos

Xiatian Zhu

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

School of Electronic Engineering and Computer Science

Queen Mary University of London

7 January 2016

In LOVING memory of my beloved grandmother.

Semantic Structure Discovery in Surveillance Videos

Xiatian Zhu

Abstract

For automatically processing and interpreting the enormous amount of video data generated by the rapid expansion of surveillance cameras, developing autonomous vision systems is essential. One generic mechanism for automated visual content analysis is to *discover and understand the intrinsic meaningful data structures*. Nonetheless, semantic structure discovery for large scale surveillance video data remains challenging due to the inherent visual ambiguity and uncertainty, potentially unreliable high-dimensional feature representations with noisy and irrelevant data, or large and unknown cross-camera variations in viewing conditions. This thesis proposes approaches to several critical video surveillance problems by deriving advanced machine learning algorithms for more accurately quantifying and mining the underlying data structure semantics. More specifically, this thesis investigates and has developed new methods for addressing four different problems as follows:

Chapter 3 The first problem is *unsupervised visual data structure discovery*, i.e. estimating the underlying data group memberships from visual observations. This is inherently challenging as visual signals can be inevitably ambiguous/noisy, e.g. due to uncontrollable variation sources like illumination and background clutter, particularly so on typical surveillance videos. Moreover, visual features are often high-dimensional, with many but unknown less-reliable feature data. To that end, this thesis proposes to identify and explore discriminative features rather than the whole feature space when measuring pairwise relationships between noisy data samples for accurately uncovering the semantic data neighbourhood structures. Specifically, a random forest based data similarity inference framework is designed, characterised by accumulating weak and subtle similarity over informative feature subspaces. This method can be utilised along with a graph based clustering algorithm for clustering visual data.

Chapter 4 The second problem is *semi-supervised visual data structure discovery* where pairwise constraints/relationships over data samples (i.e. must-link, cannot-link) are accessible. It is non-trivial to exploit pairwise constraints for helping the disclosure of meaningful data structure. This is because (1) often sparse constraints are available, thus providing only very limited information; (2) constraints are not necessarily accurate, hence misleading guidance may be imposed onto the discovery process if blindly trusting them all. In this thesis, a Constraint Propagation Clustering Random Forest model is formulated specially to leverage sparse pairwise links for more reliably measuring pairwise similarities between data pairs either constrained a priori or not. Moreover, this semi-supervised model is also characterised with favourable robustness against invalid pairwise constraints.

Chapter 5 The third one is *multi-source video data structure discovery*, significantly different from the above single-source cases. Specifically, semantic video structure analysis is investigated given heterogeneous visual and non-visual source data. Inherently, it is challenging to jointly learn such multi-source data which significantly differ in representation, scale and covariance,

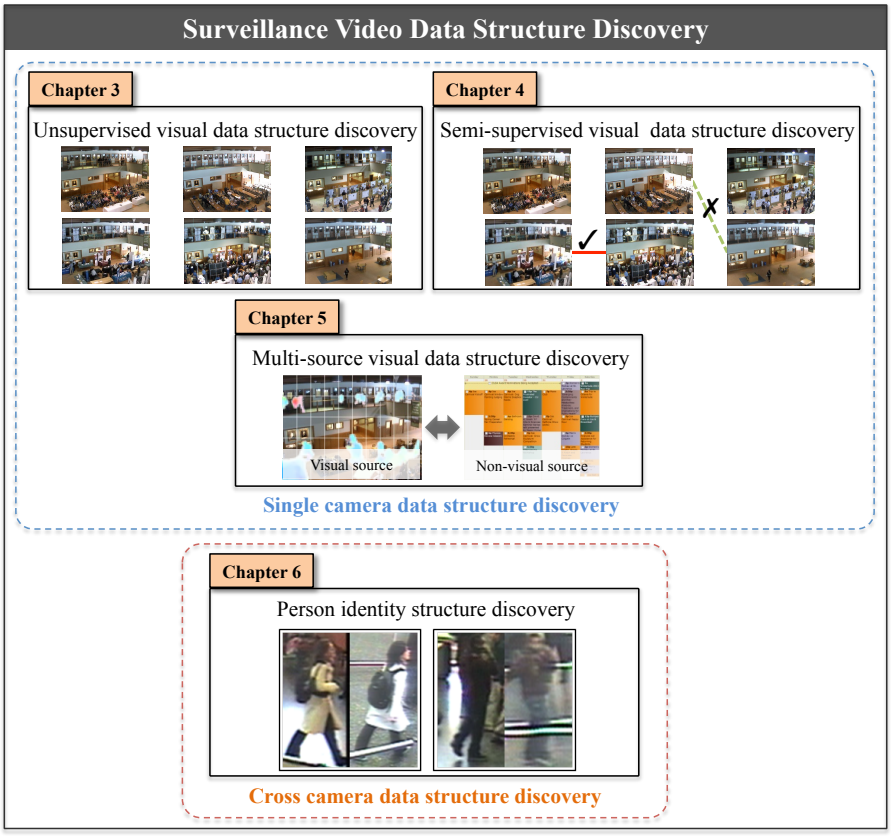


Figure 1: An overview of the main studies carried out in this thesis. According to video data source (e.g. camera) setting, all studies are grouped into two categories: single view data structure discovery (Chapters 3, 4, 5), and multi-camera data structure discovery (Chapter 6), with each chapter corresponding to a specific type of visual data structure discovery.

let alone when both visual and non-visual data in isolation can be inaccurate or incomplete. To overcome the challenges, this thesis formulates a Multi-Source Clustering Forest capable of correlating visual data and independent non-visual auxiliary information to better describe the underlying relationships among data and then facilitate video cluster revelation. The discovered clusters can be exploited to precisely summarise subtle physical events in complex scenes.

Chapter 6 The last problem is to *discover person identity structure* distributed across non-overlapping camera views, also called person re-identification (ReID). Visual data are drawn from multiple camera views, *versus* single-camera data involved in the above three problems. Therefore, visual ambiguity may be significant because of cross-view illumination variations, viewpoint differences, cluttered background and inter-object occlusions. Different from most existing appearance based models wherein ReID is achieved by matching single or multiple person images, the proposed Discriminative Video Ranking method is unique in learning a robust space-time ReID model instead from person image sequences of arbitrary starting/ending frame, random length, and unknown background clutter and occlusion. Moreover, the joint learning of both spatial appearance and space-time features in this model demonstrates significant advantages over existing methods in ReID.

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

Declaration

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

Chapter 3

1. X. Zhu, C.C. Loy and S. Gong. *Constructing Robust Affinity Graphs for Spectral Clustering*. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, June 2014. **(CVPR)**

Chapter 4

1. X. Zhu, C.C. Loy and S. Gong. *Constrained Clustering with Imperfect Oracles*. IEEE Transactions on Neural Networks and Learning Systems, 2015. **(TNNLS)**
2. X. Zhu, C.C. Loy and S. Gong. *Constrained Clustering: Effective Constraint Propagation with Imperfect Oracles*. In Proc. IEEE International Conference on Data Mining, Dallas, Texas, USA, December 2013. **(ICDM)**

Chapter 5

1. X. Zhu, C.C. Loy and S. Gong. *Learning from Multiple Sources for Video Summarisation*. International Journal of Computer Vision, 2015. **(IJCV)**
2. X. Zhu, C.C. Loy and S. Gong. *Video Synopsis by Heterogeneous Multi-Source Correlation*. In Proc. IEEE International Conference on Computer Vision, Sydney, Australia, December 2013. **(ICCV)**

Chapter 6

1. T. Wang, S. Gong, X. Zhu and S. Wang. *Person Re-Identification by Discriminative Selection in Video Ranking*. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, under minor revision. **(TPAMI)**
2. T. Wang, S. Gong, X. Zhu and S. Wang. *Person Re-Identification by Video Ranking*. In Proc. European Conference on Computer Vision, Zurich, Switzerland, September 2014. **(ECCV)**

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Prof. Shaogang Gong for his perpetual patience, continued encouragement and enthusiastic supervision. Meanwhile, I convey my special thanks to Dr. Chen Change Loy for his excellent guidance and productive support and invaluable advices. I am grateful for the support from my second-supervisor Dr. Tao Xiang. It is through their supervision and guidance that I gradually learn to independently conduct research study and creatively drive research ideas.

I would like to thank Dr. Pengwei Hao for being my independent accessor throughout my PhD study. My warm appreciation goes to all members and visiting researchers at Vision Group for their friendship and support (mostly in chronological order): Tim Hospedales, Yi-Zhe Song, Miles Hansard, Lukasz Zalewski, Yogesh Raja, Ke Chen, Parthipan Siva, Tom Haines, Zhiyuan Shi, Yanwei Fu, Ryan Layne, Howard Williams, Yi Li, Xun Xu, Wenzhao Li, Hanxiao Wang, Yongxin Yang, Zhenyong Fu, Ioannis Alexiou, Elyor Kodirov, Li Zhang, Jingya Wang, Qian Yu, Kunkun Pang, Yaowei Wang, Chunxiao Liu, Ral Martn, Brais Cancela, Shuxin Ouyang, Xiangyu Kong, Lu Tian, Lourdes Agapito, Chris Russell, Anastasios Roussos, Sara Vicente, Ravi Garg, Joao Fayad, Nikolaos Pitelis, Rui Yu, Tsz-Kin Hon, Fabio Poiesi, Riccardo Mazzon, Yiming Wang. I give big thanks to my friends at QMUL, Peng Lin, Heng Yang, Yun Zhou, Junfei Luo, Hong Liu, Mingying Song, Nian Wang, Jian Wang, Jiandong Lang, and many others for their consistent and warm encouragement. I am very grateful to my friends from Tsinghua University, Taiqing Wang and Xiaolong Ma, for our friendship and fabulous collaboration. I give sincere thanks to all friendly QMUL administrative and system support staff for their great help.

I am indebted to my family members, in particular my parents, my parents-in-law, my sisters, and my maternal grandmother for their enduring love and endless support. Most importantly, I have to thank my loving wife Ke Xu for her love and devotion all the time.

Contents

List of Figures

Nomenclature

$\mathbf{x}$	Visual data feature vector
X	Set of visual data $\mathbf{x}$
d	Number of visual feature dimension
n	Number of visual data $\mathbf{x}$
$\mathbf{y}$	Non-visual data feature vector
Y	Set of non-visual data $\mathbf{y}$
$\mathbf{z}$	Visual data feature vector after some transformation
Z	Set of visual data $\mathbf{z}$
$\mathbf{A}$	Similarity or affinity matrix

Major notations for random forest

r	Decision tree root node
s	Decision tree split or internal node
l	Decision tree leaf node
τ	Tree number of a random forest
ψ	Entropy or impurity measure
$\Delta\psi$	Information gain

General rules for notation definition

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

Chapter 1

Introduction

In visual surveillance, a fundamental task is to make sense of the massive quantity of visual data generated by the rapid expansion of Closed-Circuit TeleVision (CCTV) surveillance cameras for gaining perceptual and situational awareness of visual sensor data. This needs to extract compact, rich and expressive descriptions for video data. To that end, one common mechanism is *to identity the inherent structure underlying in large scale visual data* through measuring the similarity relationships among data samples or constructing data neighbourhoods. Whilst solving this fundamental problem is inherently challenging due to the large and hard-to-bridge semantic gap between high-level human perceptions and low-level imaginary pixel data, it can potentially benefit a variety of applications in computer vision, data mining and machine learning.

1.1 Surveillance Video Data Structure Discovery

Video surveillance is considered as one of the most important offerings in the surveillance industry. However, given the extremely enormous amount of video data produced by the incrementally growing number of surveillance CCTV cameras[1], manual data processing by human operators is prohibitively expensive and not scalable. In particular, operators involved in intensive forensic analysis of visual data from large scale multi-camera networks often face many practical challenges, including (a) data overload from extensive number of cameras, e.g. each operator may be

[1] In July 2013, it was disclosed by the British Security Industry Association that there were up to 5.9 million surveillance cameras, or averagely one camera for every 11 people in Britain (`http://www.securitynewsdesk.com/bsia-attempts-to-clarify-question-of-how-many-cctv-cameras-in-the-uk/`).

Figure 1.1: Typical surveillance control rooms with a high number of cameras. One operator may be required to monitor tens or over a hundred of cameras, e.g. by quick snapshots of all assigned screens. This shall be largely beyond the limit of human capability. Many cameras are ignored and left unmonitored for long periods of time. As a result, the surveillance system is primarily used in reactive mode, i.e. operators direct their surveillance based on receiving intelligence from external agencies, or utilise recorded footages to seek and view incidents retrospectively (Spriggs et al., 2005).

assigned tens of cameras or even more to monitor simultaneously (Spriggs et al., 2005), largely beyond their capability (see examples in Figure 1.1); (b) inherently limited attention span (Green, 1999); (c) the limited ability or difficulty to exploit other auxiliary non-visual data sources and mine informative knowledge among 'big data' for assisting the task performing process. On the other hand, the trend of rising security concerns and crime rates is increasing globally. Security is essential to the society whilst the potential public threat is everywhere, ranging from commercial and industrial scenes, to residential and other public places. In such context, the demand of intelligent video surveillance systems is surging. The ultimate aim is to facilitate the reduction of human intervention, and help process large scale surveillance video data, and achieve less threat to public safety and security.

Substantial efforts have been made towards developing automated video surveillance systems and more endeavour is expected in both the academic and industrial communities. Recently, MarketsandMarkets (MarketsandMarkets, 2014) reported that the global video surveillance market will reach $42.06 billion by 2020, growing at a CAGR (Compound Annual Growth Rate) of 16.97%. In this field, the major market players include Axis Communications AB from Sweden,

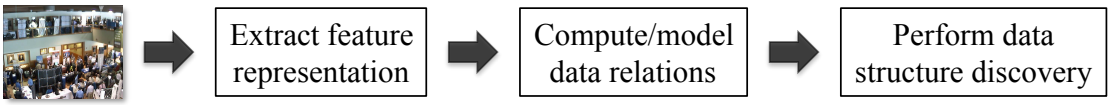


Figure 1.2: Pipeline of visual data structure discovery. Data structures refer to clusters or groups in this thesis. Clusters are task-dependent and thus may correspond to different specific concepts in distinct vision tasks. For example, a cluster may refer to a group event or an individual person.

Avigilon Corporation from Canada, Bosch Security System, Inc., Pelco by Schneider Electric from U.S., and so forth. Nevertheless, the current state-of-the-art video surveillance systems are still far from satisfactory. In particular, most existing technologies for surveillance video analysis typically depend on visual observation alone. Such systems often suffer considerably from less meaningful video abstraction and interpretation due to the large semantic gap challenge. Whilst fully unsupervised or semi-supervised analysis of visual data captured from public spaces is often challenged by the inherent ambiguities and uncertainties of visual appearance. This can be due to the large and unknown variations in lighting, image quality, imaging noises, diversity of object pose and appearance, and severe random occlusions in crowded scenes. As a result, existing systems that rely on the whole and potentially noisy visual features are likely to produce sub-optimal results. In addition, many contemporary techniques are limited by design in exploiting sufficiently the available visual data. To solve the aforementioned problems demands the innovation of more robust and advanced computer vision algorithms and surveillance systems.

1.1.1 Definition of Data Structure Discovery

Abstractly, the essence of many video surveillance systems is to acquire compact and meaningful explanations and/or descriptions for the visual data under consideration. The visual data to be processed is often of large scale and unstructured. Therefore, one typical and natural processing mechanism is data structure analysis and discovery. Specifically, an intelligent vision system takes three steps for discovering the latent data structure (Figure 1.2):

1. Extracting visual feature representations for data samples, e.g. colour, texture, optic flow, object detections.
2. Modelling and computing the quantitative relationships between samples, e.g. data pairwise distance/similarity. A model training stage for parameter optimisation may or may not be required, depending on the specific learning strategy.
3. Performing visual data structure discovery for providing task-specific interpretation based

on the learned model and/or the inferred numerical measures.

Among this process, one key issue is how to model and compute the numerical relations between data. This is the focus of the works presented in this thesis. Formally, *data structures are defined as the cluster or group memberships over a collection of data* in this context. Clusters are task-dependent and should respect human perception, e.g. a cluster may refer to a certain physical activity/event (Figure 3.2c).

Automatic cluster structure discovery is an essential means of surveillance video analysis since it provides a concise and manageable description and index for overwhelmingly large video data. Several important uses of discovered clusters include (1) facilitating video management; (2) providing easy browsing capability; (3) allowing efficient exploration of a video dataset; (4) summarising video data. All these functionalities are crucial for video surveillance.

1.1.2 Definition of Problems

In typical video surveillance, visual data under consideration are usually drawn from a single camera view or multiple camera views (e.g. multi-camera CCTV networks). From this data source perspective, the problems/tasks considered in this thesis are categorised into two classes:

1. **Single-camera visual data structure discovery**. The aim is to identify the *inherent cluster or group structures* of the given single-camera visual data, e.g. group events. Particularly, three specific problems are involved as below.
   (a) *Unsupervised visual data structure discovery*: The typical unsupervised cluster analysis (or data clustering) is considered, wherein only data features are available without the presence of any other knowledge.
   (b) *Semi-supervised visual data structure discovery*: The semi-supervised data clustering problem is studied in case that a number of pairwise constraints are accessible apart from visual data features.
   (c) *Multi-source data structure discovery*: The multi-source data clustering problem is investigated, assuming the availability of additional auxiliary non-visual data sources, e.g. a single data sample is associated with multiple information components. The discovered structures can be further exploited for video summarisation.
2. **Cross-camera visual data structure discovery**. Multiple cameras allow video surveillance to monitor a wide area, physically extending the limited viewing area by single cam-

eras. For distributed multi-camera systems, an essential task is to associate people across camera views at different locations with no overlapping field of view. This is also known as the *person re-identification* (ReID) problem. In other words, the notion of cluster in this cross-view person ReID scenario is specifically defined as person identity, and one aims to build clusters each corresponding to one particular person of interest. This is a finer-grained data structure when compared to the coarse group activity/event cluster. Person ReID enables to discover and reason about the latent person-specific structured activities taking place over extended public spaces, facilitating individual global behaviour analysis beyond single localised camera views.

Figure 1 (in Abstract) summarises and organises together all these problems considered and investigated for obtaining a global understanding of the whole thesis.

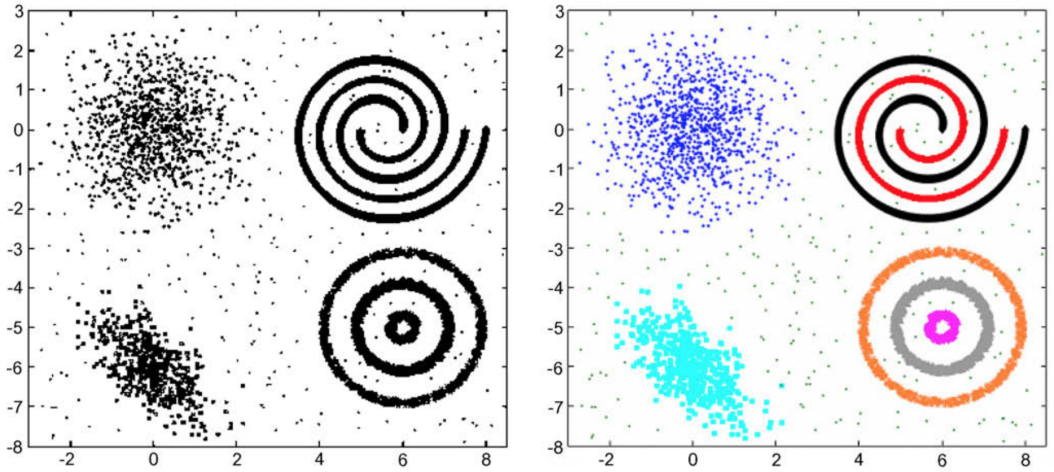
1.2 Challenges, Hypotheses, and Solutions

This section discusses the approaches/solutions proposed to solve these problems (Section 1.1.2), together with particular challenges in addressing the problem, and the hypothesis behind the presented solutions.

1.2.1 Unsupervised Visual Data Structure Discovery

Unsupervised visual data structure discovery or clustering is a fundamental and popular approach to video data understanding, essential for many computer vision tasks, including supervised learning like general classification and regression problems. Also, it promises immense potential for a wide range of applications in data mining and pattern recognition (Jain, 2010).

Challenges Performing unsupervised visual data clustering is intrinsically challenging especially given complex data that are often of high dimension and represented by less reliable features, whilst no additional prior knowledge or supervision is available for helping resolve uncertainty. Trusting all available visual features blindly for measuring data pairwise distance and similarity is susceptible to unreliable and/or noisy features, particularly so for real-world visual data, e.g. images and videos where signals can be inevitably inaccurate and unstable owing to uncontrollable sources of variation, changes in illumination, context, occlusion and background clutters (Gong et al., 2011). Moreover, confining the notion of localised data pairwise distance to the L_2-norm metric implicitly imposes unrealistic assumption on complex data structures that



(a) Input data (b) Desired clustering

Figure 1.3: An example from (Jain, 2010) showing the challenges involved in cluster analysis. None of the existing clustering algorithms can discover accurately all these clusters, which however are very apparent to human analyst.

do not necessarily possess the Euclidean behaviour. How to learn a semantically meaningful distance and similarity metric remains open. Figure 1.3 shows an example for illustrating the challenges in cluster analysis.

Hypothesis Given high-dimensional and possibly noisy visual data feature spaces, it is hypothesised that underlying data cluster structures can be inferred and discovered more accurately over discriminative feature subspaces, rather than using the entire feature space (Figure 1.4).

Solution The goal is to infer accurate pairwise similarity between visual samples so as to construct more meaningful affinity graphs for facilitating unsupervised data cluster discovery using existing graph based algorithms, e.g. spectral clustering. Instead of considering the complete feature space as a whole, the proposed model is designed to avoid less informative visual features by measuring inter-sample proximity via discriminative feature subspaces, yielding similarity graphs that better express the underlying structures in visual data. Moreover, the Euclidean assumption is relaxed for data similarity inference by following the information-theoretic definition of data similarity presented in (Lin, 1998), which states that different similarities can be induced from a given sample pair if distinct propositions are taken or different questions are asked about data commonalities. Motivated by a similar idea, the proposed model derives pairwise similarities of arbitrary sample pairs from an exhaustive set of comparative tests, using different feature variables with distinct inherent semantics as criteria. Such subtle similarities distributed over discriminative visual feature subspaces are combined automatically and effectively for produc-

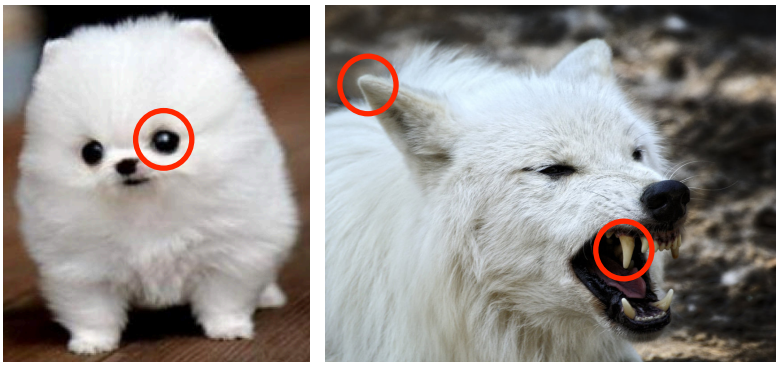


Figure 1.4: An example for demonstrating the importance of feature selection. Whilst their holistic appearances share much similarity, the significant discriminations are hidden in some particular regions as indicated by red circles. This necessitates selective matching other than global comparison in the data visual feature space.

ing robust pairwise affinity matrices.

1.2.2 Semi-Supervised Visual Data Structure Discovery

The ill-definition of unsupervised data structure discovery inspires the search and exploration of other information sources. In some circumstances one may have access to prior belief that pairs of samples should or should not be assigned with the same cluster or high-level explanation. Exploiting this prior belief as additional constraint information or weak supervision to influence the cluster discovery process can obtain a data group structure more closely resembling human perception. Such constraints are often available in small quantity, and expressed in the form of pairwise link, namely must-link - a pair of samples must be in the same cluster, and cannot-link - a pair of samples belong to different clusters. Clustering with such external information is also known as *constrained clustering* (Wagstaff et al., 2001; Basu et al., 2004a, 2008) or *semi-supervised clustering* (Basu et al., 2002, 2004b; Kulis et al., 2009; Araujo, 2015) due to its commonality with semi-supervised learning (Blum and Mitchell, 1998; Zhu et al., 2003; Chapelle et al., 2006) where supervision is provided only over a limited number of data, i.e. a number of sample pairs are connected with either must-link or cannot-link whilst the others are not. In the context of data cluster analysis, pairwise constraints are more natural than commonly-used class labels because cluster membership is the learning target. Also, the annotation for the latter requires human annotators to own more knowledge and thus more demanding. The *objective* is to exploit this small amount of pairwise supervision effectively to help reveal the visual data partitions/groups that capture consistent concepts as perceived by human or as indicated by the pairwise constraints.

Challenges Two important but non-trivial questions remain unsolved despite extensive research effort has been expended on the constrained clustering problem in the last decade:

1. *Sparse constraint propagation* Whilst constraints can be readily transformed into pairwise similarity measures, e.g. assign 1 to the similarity between two must-linked samples, and 0 to that between two cannot-linked samples (Kamvar et al., 2003), samples labelled with link preference are typically insufficient since exhaustive pairwise labelling is laborious and tedious. As a results, the limited number of constraint samples are usually employed together with data features to positively affect the similarity measures over unconstrained sample pairs so that the yielded similarities are closer to the intrinsic structures in data. Such a similarity distortion/adaptation process is often known as *constraint propagation* (Lu and Carreira-Perpinán, 2008; Lu and Ip, 2010). Effective constraint propagation relies on robust identification of unlabelled nearest neighbours (NN) around the labelled samples in the feature space. Often, the NN search is susceptible to noisy or ambiguous features, especially so on image and video datasets. Trusting all the available features blindly for NN search (as what most existing constrained clustering approaches (Wagstaff et al., 2001; Lu and Carreira-Perpinán, 2008; Lu and Ip, 2010) did) is likely to result in suboptimal constraint diffusion. It is challenging to determine how to propagate their influence effectively to neighbouring unlabelled points. In particular, it is non-trivial to reliably identify the neighbouring unlabelled points for propagation.
2. *Noisy constraints from imperfect oracles* Human annotators (oracles) may provide invalid/mistaken constraints. For instance, a portion of the 'must-links' are actually 'cannot-links' and vice versa. For example, annotations or constraints obtained from online crowdsourcing services, e.g. Amazon Mechanical Turk (Kittur et al., 2008), are very likely to contain errors or noises due to data ambiguity, unintentional human mistakes or even intentional errors by malicious workers (Kittur et al., 2008; Patterson and Hays, 2012). Learning such constraints blindly may result in sub-optimal cluster formation. Most existing methods make an unrealistic assumption that constraints are acquired from perfect oracles thus they are noise-free. It is non-trivial to quantify and determine which constraints are noisy prior to knowing their true cluster memberships.

Hypothesis The hypotheses to validate in this study are:

1. Pairwise constraints can be propagated more effectively and accurately to unconstrained

samples through nearest neighbours defined over discriminative feature subspaces, rather than the whole feature space.

2. As propagating noisy pairwise links may lead to suboptimal clustering results, it is essential to discover and measure the noisy degree of individual constraints by measuring the relationships between data features and the given links for maximising the benefits of potentially inaccurate constraints from imperfect oracles.

Solution To address the sparse and noisy pairwise constraint issues in semi-supervised visual data structure discovery, this thesis formulates a COnstraint Propagation Random Forest (COP-RF), not only capable of effectively propagating sparse pairwise constraints, but also able to deal with noisy constraints produced by imperfect oracles. The COP-RF is flexible in that it generates an affinity matrix that encodes the constraint information for existing spectral clustering methods (Ng et al., 2002; Zelnik-manor and Perona, 2004; Von Luxburg, 2007; Xiang and Gong, 2008) or other pairwise similarity based clustering algorithms for constrained clustering.

More precisely, the proposed model allows for effective sparse constraint propagation through using the NN samples that are found in discriminative feature subspaces, rather than those found considering the whole feature space, which can be suboptimal due to noisy and ambiguous features as shown in Figure 1.4. This is made possible by introducing a new objective split function into COP-RF, which searches for discriminative features that induce the best data subspaces while simultaneously considering the model parameters that best satisfy the data-level constraints imposed. To identify and filter noisy constraints generated from imperfect oracles, a constraint inconsistency quantification algorithm based on the outlier detection mechanism of random forest is introduced. Figure 4.1 shows an example of illustrating how a COP-RF is capable of discovering data partitions close to the ground truth cluster structures despite that it is provided only with sparse and noisy pairwise constraints.

1.2.3 Multi-Source Data Structure Discovery

In the context of video surveillance, there may exist a number of non-visual auxiliary information. Examples of non-visual sources include weather report, GPS-based traffic data, geo-location data, textual data from social networks, and on-line event schedules (Figure 1.5). The auxiliary data sources are beneficial to visual data modelling because despite that visual and non-visual data may have very different characteristics and are of different natures, they depict the common

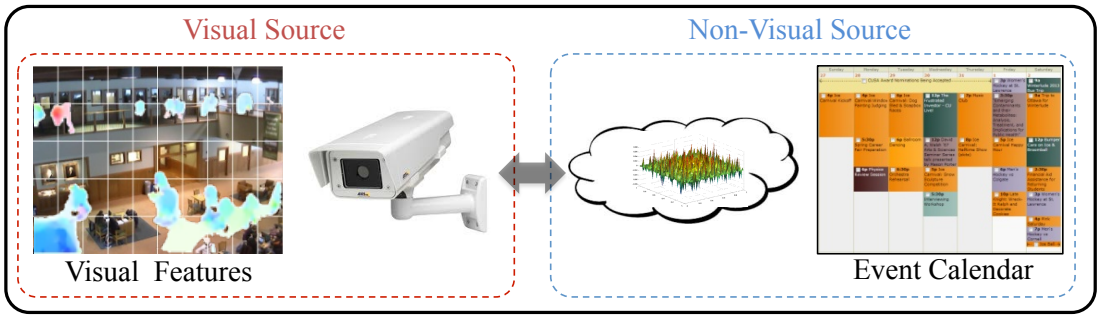


Figure 1.5: A multi-source data example in visual surveillance. Beyond the visual source from the camera, some non-visual data sources may be accessible and helpful in video analysis.

physical phenomenon in a scene. So, they are intrinsically correlated although may be mostly indirect in some latent spaces.

Challenges Nevertheless, it is non-trivial to formulate a framework that exploits both visual and non-visual data for video content analysis, both algorithmically and in practice.

Algorithmically, unsupervised mining of latent correlations and interactions between heterogeneous data sources faces a number of challenges: (1) Disparate sources significantly differ in representation (continuous or categorical), and largely vary in scale and covariance. This is also known as the heteroscedasticity problem (Duin and Loog, 2004). In addition, the dimension of visual sources often exceeds that of non-visual information to a great extent ($>$2000 visual dimensions vs. $<$10 non-visual dimensions). Owing to this dimensionality discrepancy problem, a straightforward concatenation of features will result in a representation unfavourably inclined towards the imagery data. (2) Both visual and non-visual data in isolation can be inaccurate and incomplete.

In practice, auxiliary data sources, e.g. weather, traffic reports, and event time tables, may be rather unreliable in availability. Specifically, the reports may not be released on-the-fly at a synchronised time stamp with the surveillance video stream. In addition, existing video control rooms may not necessarily have direct access to these sources. This renders models that expect complete visual and non-visual information during deployment impractical.

Hypothesis In this study, exploiting the available non-visual auxiliary information is hypothesised to positively complement the unilateral perspective from visual observations. In particular, effectively discovering and exploiting such a latent correlation space can facilitate the underlying data structure discovery and bridge the semantic gap between low-level visual features and high-level interpretation of video content.

Solution To overcome the above challenges, this study presents a unified multi-source data discovery framework capable of performing joint learning given heterogeneous multi-sources (Figure 5.1). The visual data is considered as the *main source* and non-visual data as the *auxiliary sources*, since visual information still plays the main role in video content analysis. More specifically,

1. During training, the access to both visual and non-visual data is assumed. The proposed model performs multi-source data clustering and discovers a set of visual clusters tagged along with non-visual data distribution, e.g. different weathers and traffic speeds. The model is termed as *multi-source model*.
2. During the deployment stage, only the availability of previously-unseen video data is assumed since non-visual data may not be accessible due to the aforementioned limitations.

This learned multi-source learning model can be applied for video summarisation, conventionally based on visual feature analysis and object detection or segmentation alone. Specifically, as the learned model has already captured the latent structure of heterogeneous types of data sources, the model can be used for semantic video clustering and non-visual tag inference on previously-unseen video sequence, even without the non-visual data. Subsequently, key clips are automatically selected from the discovered clusters. The final summary video can be produced by chronologically compositing these key clips enriched by the inferred tags.

1.2.4 Person Identity Structure Discovery

Person identity distribution structure discovery is typically realised by person re-identification (ReID). The state-of-the-art person re-identification methods perform cross view people association mostly by matching spatial appearance features (e.g. colour and intensity gradient histograms) using a pair of single-shot person images (Hirzer et al., 2012; Farenzena et al., 2010; Prosser et al., 2010; Zhao et al., 2013b). However, single-shot appearance features of person are intrinsically limited due to the inherent visual ambiguity caused by clothing similarity among people in public spaces, appearance changes from cross-view illumination variations, viewpoint differences, cluttered background and inter-object occlusions. Exploiting other information cues are required for improving the performance of current person re-identification systems. In practical surveillance settings, continuous video data are often accessible beyond discrete individual person images. It is desirable to explore space-time information from the available videos or

(a) Cross-view lighting variations (b) Camera viewpoint changes

(c) Clothing similarity (d) Background clutter and occlusions

Figure 1.6: Person re-identification challenges in public space scenes (UK , 2008). (a,b): The two images in each bounding box refer to the same person observed in different cameras.

image sequences of people for assisting re-identification in public spaces.

Challenges It is inherently non-trivial to extract reliably person-discriminative space-time information from image sequences, especially when the videos are captured at crowded public scenes. This is because:

1. The starting/ending frames of each sequence may correspond to arbitrary walking phases and thus two compared sequences are mostly unaligned.
2. Sequences of pedestrians have varying number of walking cycles and a holistic matching between sequences may yield suboptimal match in parts of the sequences.
3. Image sequences of people walking in public spaces consist of missing or corrupted frames due to background clutter and random inter-object occlusions (see examples in Figure 1.6). This makes walking phase detection unreliable.

Hypothesis For re-identifying people using largely unaligned and noisy person videos, it is hypothesised that the key to learn an effective image sequence ReID model can lie in selecting and exploiting informative and discriminative video fragments from the whole sequences for facilitating the extraction of reliable identity-discriminative space-time dynamic visual features.

Solution The aim of this study is to construct a discriminative video matching framework for person ReID by selecting more reliable space-time features from person videos, beyond the often-adopted spatial appearance features. To that end, it assumes the availability of image sequences of people which may be highly noisy, i.e., with arbitrary sequence duration and starting/ending

frames, unknown camera viewpoint/lighting variations during each image sequence, also with likely incomplete frames due to uncontrolled occlusions. These videos are called *unregulated* image sequences of people (Figure 1.6 and Figure 6.4). More specifically, an approach to Discriminative Video fragments selection and Ranking (DVR) is proposed, based on a robust space-time and appearance feature representation given unregulated image sequences of people.

1.3 Contributions

The contributions made in this thesis are summarised below:

1. Chapter 3: A unified and generalised visual data similarity inference framework is formulated based on the unsupervised clustering random forest for more accurate unsupervised data structure discovery. The pairwise affinity matrix generated by the proposed model automatically possesses the local neighbourhood. Thus, no additional Gaussian kernel is needed to enforce locality.
2. Chapter 4: A discriminative-feature driven semi-supervised visual data structure discovery approach is formulated for effective sparse constraint propagation. Existing methods fundamentally ignore the role of feature selection in this problem. Further, a new method is presented to cope with potentially noisy constraints based on constraint inconsistency measures, a problem that is largely unaddressed by existing constrained clustering algorithms. The sparse and noisy constraint issues in constrained clustering are inextricably linked but no existing constrained clustering method addresses them in a unified framework. To our knowledge, this is the very first study that addresses them jointly.
3. Chapter 5: A multi-source data learning framework capable of discovering semantic video cluster structures using collectively heterogeneous visual and non-visual data is proposed, achieving effective multi-source data structure discovery. This is made possible by formulating a Multi-Source Clustering Forest (MSC-Forest) that seamlessly handles multi-heterogeneous data sources dissimilar in representation, distribution, and covariate. Although both visual and non-visual data in isolation can be inaccurate and incomplete, this proposed model is capable of uncovering and subsequently exploiting the shared latent correlation for better data structure discovery. This multi-source model is novel in its ability to accommodate partially or completely missing non-visual sources. In particular, a joint information gain function that is capable of dynamically adapting to arbitrary amount of

missing non-visual information during model learning is introduced.

4. Chapter 6: A multi-fragment based appearance and space-time feature representation of image sequences of people is derived for person identity structure discovery over distributed camera views. This representation is based on a combination of HOG3D and colour features and optic flow energy profile over each image sequence, designed to break down automatically unregulated video clips of people into multiple fragments. More importantly, a discriminative video ranking model for cross-view person re-identification by simultaneously selecting and matching more reliable space-time features from video fragments is developed. The model is formulated using a multi-instance ranking strategy for learning from pairs of image sequences drawn from non-overlapping camera views. The proposed method can significantly relax the strict assumptions made by gait recognition techniques.

1.4 Thesis Outline

This thesis is organised as follows, with all chapters structured as shown in Figure 1.7.

Chapter 2 presents a review on various existing data structure discovering strategies and approaches, related learning models such as random forests, and video surveillance applications like video summarisation.

Chapter 3 explains a unified framework for unsupervised visual data structure discovery e.g. clustering. Specifically, the chapter describes a generic data clustering approach based on unsupervised random forest and spectral clustering algorithm. It is characterised by quantifying and cumulating subtle and weak localised data similarity defined over discriminative feature subspaces. Visual data clustering experiments are conducted to evaluate the advantages of the proposed algorithm by extensively comparing most contemporary methods.

Chapter 4 describes a semi-supervised visual data discovery (or constrained clustering) approach particularly designed for solving the sparse and noisy pairwise constraints issues. In particular, the chapter details a constraint propagation random forest model capable of more effectively propagating a small number of constraints, and an algorithm for pairwise constraint consistency measure. Finally, the efficacy of the proposed method is validated by extensive comparisons with state-of-the-art clustering models on various types of data, including images and videos.

Chapter 5 provides detailed modelling and application explanations on the multi-source data structure discovery. Particularly, it demonstrates that the proposed Multi-Source Clustering Forest is capable of not only effectively extracting high-level knowledge from heterogeneous multi-source data for constructing consistent visual data clusters, but also providing more accurate understanding on previously-unseen video data. This multi-source model is finally applied for video summarisation.

Chapter 6 presents an person identity structure discovery method with unconstrained image sequences, a.k.a. person re-identification. In contrast to existing algorithms that typically rely on person spatial appearance, this model is unique in the ability of extracting discriminative person-specific space-time features from even largely noisy and unaligned image sequences by automatically selecting informative video fragments for model learning. Thorough experiments are carried out to demonstrate the effectiveness of the proposed approach by extensive comparison with contemporary gait recognition, temporal sequence matching, and single-/multi-shot person re-identification methods.

Chapter 7 provides conclusion and suggests a number of research problems and directions to be pursued as further work.

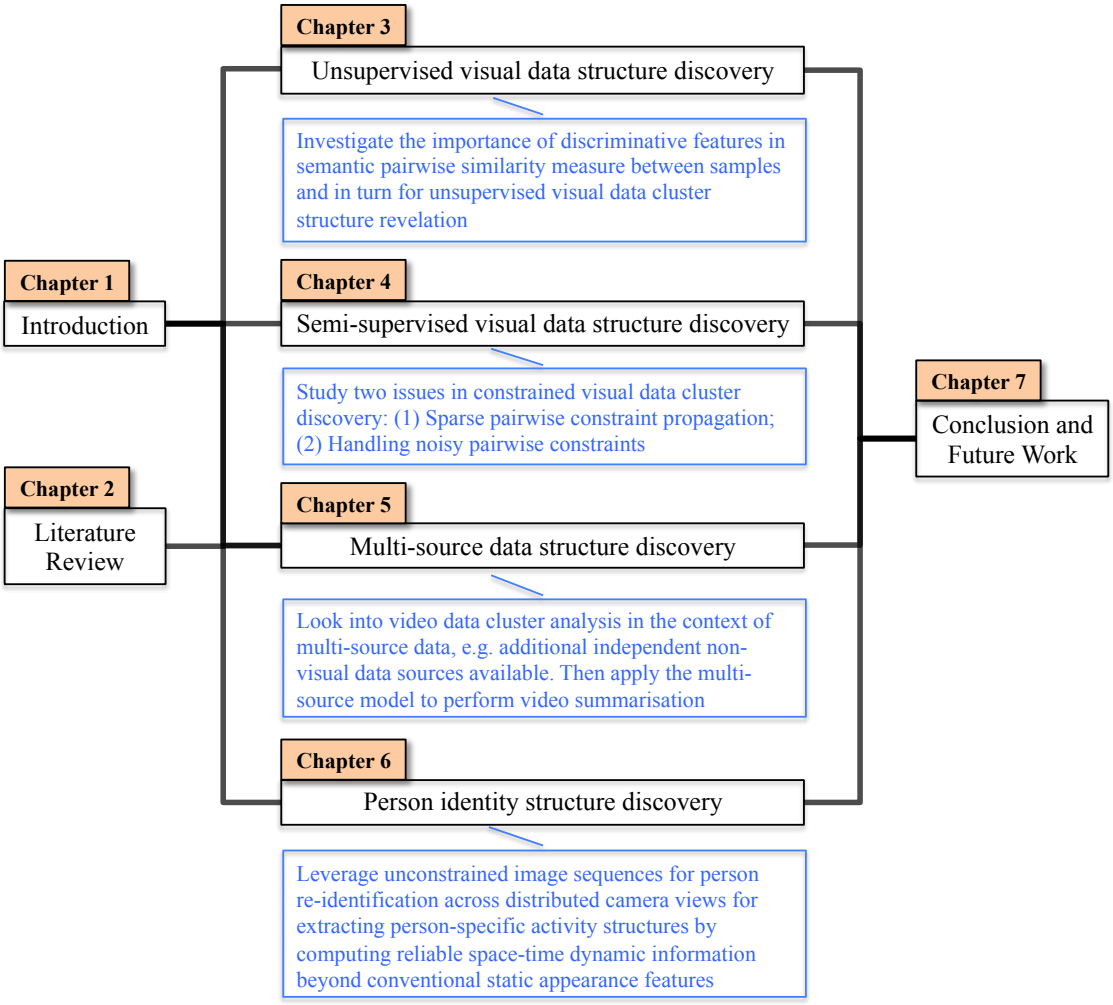


Figure 1.7: Summarisation and structure of all chapters.

Chapter 2

Literature Review

2.1 General Learning Strategies

Automatically interpreting and understanding large scale surveillance video data by developing intelligent machine vision systems remains challenging despite the significant progress made during last decades. Generally, there are three canonical learning paradigms:

1. *Supervised learning*, where some type of labels or annotations associated with data samples are provided for model learning (Figure 2.1-(a)).
2. *Unsupervised learning*, where no data annotation is accessible and the system aims to form clusters or groups (Figure 2.1-(d)). This is also known as *clustering*.
3. *Semi-supervised learning*, where data labels are partially available and the system exploits labelled and unlabelled data for model building since both types of data can provide useful information. This is a hybrid setting of supervised and unsupervised learning (Chapelle et al., 2006) (Figure 2.1-(b-c)).

Alternatively, the ultimate objective of all these paradigms can be understood and summarised as *to extract compact and semantic description/interpretation for the target data by discovering underlying meaningful structures* using some learning algorithm for reasoning the intrinsic data ambiguities and uncertainties in mutual relation, e.g. numerical distance and similarity.

In machine learning and computer vision, random forests (Breiman, 2001; Criminisi and Shotton, 2012) have a rich and successful history, while being considered to be close to an ideal learner (Friedman et al., 2001). They compare favourably with other machine learning

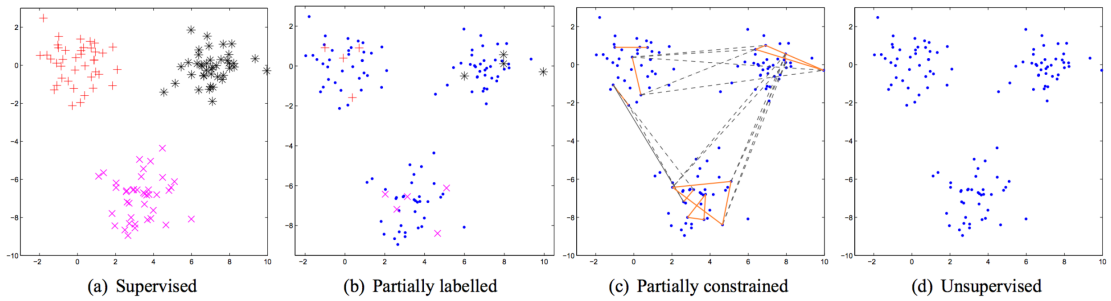


Figure 2.1: An illustration of various learning settings: dots correspond to data samples without any labels. Samples with labels are denoted by circles, asterisks and crosses. In (c), the must-link and cannot-link pairwise constraints are represented by solid and dashed lines, respectively. This figure is borrowed from (Lange et al., 2005).

algorithms and have empirically demonstrated to outperform most state-of-the-art learners particularly in high dimensional data problems (Caruana and Niculescu-Mizil, 2006; Caruana et al., 2008). Their merits include: (1) fast training and evaluation, (2) robustness to label noises, (3) inherent multi-class capability and feature selection mechanism, (4) suitability for parallel processing, and (5) promising performance for high-dimensional input data. Inspired by the great success of random forests, most of the proposed methods (e.g. in Chapters 3, 4, 5) are formulated and designed based upon them. Below, a fundamental review on random forests is firstly provided, followed by a more holistic survey about various data structure discovery studies.

2.2 Random Forests

A random forest (Breiman, 2001) is an ensemble model of multiple decision trees. Notable early decision tree models include "Classification and Regression Trees (CART)" (Breiman et al., 1984), and "C4.5" (Quinlan, 1993). These tree models were often utilised individually instead of in the ensemble form. Amit and Geman (1994, 1997) firstly constructed and exploited ensembles of trees for obtaining greater accuracy and generalisation on the problem of handwritten digits recognition. They also suggested a simple but powerful ensemble model that is still widely adopted thus far – the mean of tree-level probabilities/predictions. This modelling shares a similar principle as the well known boosting ensemble model of Schapire (1990) where iterative reweighting of training data are used to build a strong model as a linear aggregation of many weak ones. Further, Ho (1995) showed the advantages of trees trained using randomised partitioning of feature space over "C4.5" trees and similarly for forest models (Ho, 1998). The popular use of random forests can be largely attributed to the work by Breiman (2001). Importantly, the role of

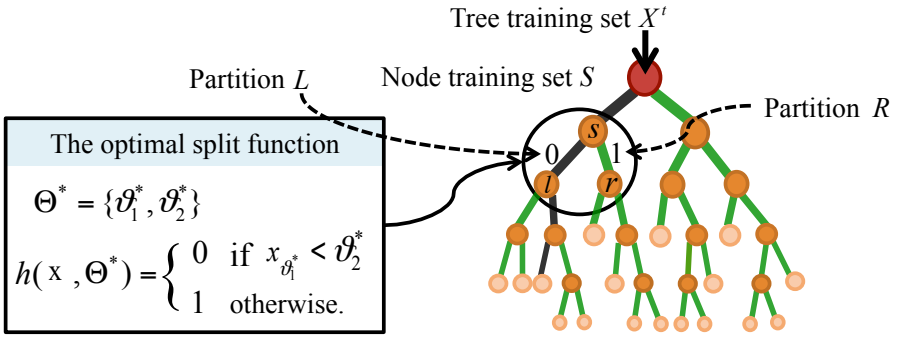


Figure 2.2: An illustrative example on the training process of a decision tree.

random forests among machine learning algorithms are further consolidated by this study. Many different random forest variants have been proposed and applied in various machine learning and computer vision problems since then, particularly in recent several years.

In general, random forests can be divided into two classes: (1) supervised forests, including classification (labels are discrete) and regression (labels are continuous) variants; (2) unsupervised forests, or clustering forests. The details of these forest models will be discussed next.

2.2.1 Classification Forests

A general form of random forests is the classification forests. A classification forest (Breiman, 2001) is an ensemble of τ_{class} binary decision trees: $F \rightarrow \mathbb{R}^k$, with F the data feature space, $\mathbb{R}^k = [0,1]^k$ denoting the space of class probability distribution over the label space $C = \{1,\ldots,k\}$ including a total of k different categories.

Tree training Decision trees are learned independently from each other, each with a random training set $X^t \subset X$, i.e. bagging (Breiman, 2001). It is this independence property that allows parallel processing, either model training or testing. Growing a decision tree involves a recursive node splitting procedure. The training of each internal (or split) node s is a process of optimising a binary split function defined as

$$\mathrm{h}(\mathbf{x}, \boldsymbol{\vartheta}) = \begin{cases} 0, & \text{if } x_{\vartheta_1} < \vartheta_2, \\ 1, & \text{otherwise.} \end{cases} \tag{2.1}$$

This split function is parameterised by two parameters: (i) a feature dimension x_{ϑ_1}, with $\vartheta_1 \in \{1,\ldots,d\}$, and (ii) a feature threshold $\vartheta_2 \in \mathbb{R}$. We denote the parameter set of the split function as $\boldsymbol{\vartheta} = [\vartheta_1, \vartheta_2]$. All arrival samples of a split node will be channelled to either the left or right child node according to the output of Equation (2.1).

The optimal split parameter $\hat{\boldsymbol{\vartheta}}$ is chosen via

$$\hat{\boldsymbol{\vartheta}} = \underset{\Theta}{\operatorname{argmax}} \Delta\psi, \tag{2.2}$$

where $\Theta = \left\{ \boldsymbol{\vartheta}^i \right\}_{i=1}^{d_{\text{try}}(|S|-1)}$ represents a parameter set over d_{try} randomly selected features, with S the sample set arriving at the split node s. The cardinality of a set is given by $|\cdot|$. Particularly, multiple candidates of data splitting are attempted on d_{try} random feature-dimensions during the above node optimisation process.

Typically, a greedy search strategy is exploited to identify $\hat{\boldsymbol{\vartheta}}$. The information gain $\Delta\psi$ is formulated as

$$\Delta\psi = \psi_s - \frac{|L|}{|S|}\psi_{lc} - \frac{|R|}{|S|}\psi_{rc}, \tag{2.3}$$

where s, lc, rc refer to a split node, the left and right child node, respectively. The sets of data routed into lc and rc are denoted as L and R, and $S = L \cup R$ as the sample set residing at s. The ψ can be computed as either (Criminisi and Shotton, 2012) the entropy or Gini impurity (Breiman et al., 1984). In this study we utilise the Gini impurity due to its simplicity and efficiency. The Gini impurity is computed as

$$\psi_{\text{gini}} = \sum_{i \neq j} p_i^{\text{prp}} \times p_j^{\text{prp}}, \tag{2.4}$$

with p_i^{prp} and p_j^{prp} being the proportion of samples belonging to the i-th and j-th category, respectively. The computational complexity of Equation (2.4) is $O(1)$, i.e. constant, because it is computed over the category distribution and therefore very efficient.

By doing so, an internal node s selects the most *discriminative* (i.e. maximising the information gain) feature from d_{try} candidates as its split variable and exploits it to partition the training data S. This process is repeated throughout the whole tree training stage until some stopping criterion is satisfied, e.g. the number of training samples S arriving at a node is equal to or smaller than a threshold ϕ. After the node splitting process stops, leaf nodes are formed. For each leaf l, a predictor model can be estimated from the labels of training samples falling into this node, e.g. a probabilistic histogram distribution over all k categories $p_{\text{post}}^l(c), c \in C$. Figure 2.2 provides an illustration on the training procedure of a decision tree.

During testing, each decision tree yields a posterior distribution $p_{\text{post}}^t(c|\mathbf{x}^*)$ for a given unseen sample $\mathbf{x}^* \in F$. This tree-level prediction is made based on the predictor model of the leaf where $\mathbf{x}^*$ falls into. The output probability of forest is obtained via averaging as

$$p_{\text{post}}(c|\mathbf{x}^*) = \frac{1}{\tau_{\text{class}}} \sum_{t=1}^{\tau_{\text{class}}} p_{\text{post}}^t(c|\mathbf{x}^*). \tag{2.5}$$

The final class label is obtained as $\hat{c} = \text{argmax}_{c \in C} p(c|\mathbf{x}^*)$.

2.2.2 Regression Forests

Similar to classification forests, a regression forest (Breiman, 2001) is a collection of τ_{reg} regression trees: $F \rightarrow \mathbb{R}$, The main difference is that label data $\tilde{c} \in \mathbb{R}$ are continuous, compared to the discrete labels in classification forests. The training is performed with the same process, e.g. using Equations (2.1), (2.2) and (2.3). Due to the continuous nature of the labels, a different way is needed for computing information gain. One common metric to measure information or entropy is least squares regression (Breiman et al., 1984). Formally, the regression impurity over a training set S is computed as

$$\psi_{\text{lsr}} = \frac{1}{|S|} \sum_{i=1}^{|S|} (\tilde{c}_i - \frac{1}{|S|} \sum_{i=1}^{|S|} \tilde{c}_i)^2, \tag{2.6}$$

where $\tilde{c}_i$ represents the label value of the i-th training sample $\mathbf{x}_i \in S$. The complexity of Equation (2.6) is linear with the number of samples n, e.g. $O(n)$.

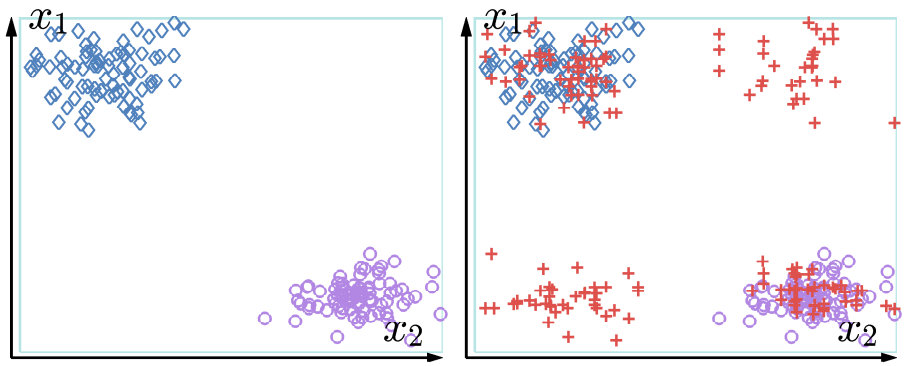
For leaf predictor models $p^l_{\text{post}}(\mathbf{x})$, there are multiple alternatives, such as constant, linear or polynomial models (Criminisi and Shotton, 2012).

In testing, the final forest prediction of a previously unseen sample $\mathbf{x}^*$ is the average of all regression tree outputs $p^t_{\text{post}}(\mathbf{x}^*)$:

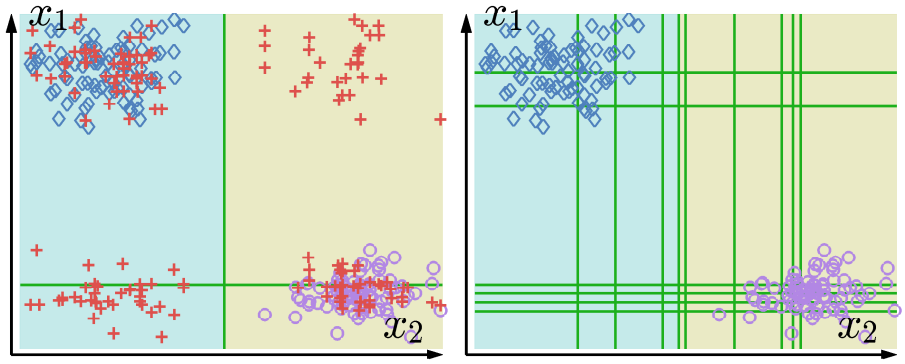
$$p_{\text{post}}(\mathbf{x}^*) = \frac{1}{\tau_{\text{reg}}} \sum_{t=1}^{\tau_{\text{reg}}} p^t_{\text{post}}(\mathbf{x}^*). \tag{2.7}$$

2.2.3 Clustering Forests

In contrast to classification and regression forests, clustering forests (Breiman, 2001) require no ground truth label information during the training phase. A clustering forest consists of τ_{clust} binary decision trees. The leaf nodes in each tree define a spatial partitioning of the training data. Several unsupervised splitting strategies have been proposed (Breiman, 2001; Yu et al., 2011; Criminisi and Shotton, 2012; Pei et al., 2013). By adopting the pseudo two-class algorithm (Breiman, 2001; Shi and Horvath, 2006), the training of a clustering forest can be performed using the classification forest optimisation approach. Specifically, we add n pseudo samples $\bar{\mathbf{x}} = \{\bar{x}_1, \ldots, \bar{x}_d\}$ (Figure 2.3-b) into the original data space X (Figure 2.3-a), with $\bar{x}_i \sim \text{pdf}(x_i)$ sampled from certain probability distribution $\text{pdf}(x_i)$. With this data augmentation strategy, the clustering problem becomes a canonical classification problem that can be solved by the classification forest training method as discussed above. The key idea behind this algorithm is to



(a) Original data (b) Added pseudo data



(c) First 2 decisions (d) Resulting partitions

Figure 2.3: An illustration of performing data partition with a random forest over a toy dataset. Original toy data samples (a) are labelled as class 1, whilst the red-coloured pseudo-points '+' (b) as class 2. A random forest performs a two-class classification on the augmented space (c). (d) The resulting data partitions on the original data.

partition the augmented data space into dense and sparse regions (Figure 2.3-c,d) (Liu et al., 2000). In our proposed models, we adopt this strategy because of its simplicity and efficiency (with $Q(1)$ computational complexity as Equation (2.3)) thus scalable. We utilise the empirical marginal distributions of the feature variables owing to its favourable performance and low computational cost (Shi and Horvath, 2006).

A second strategy assumes that two features with a larger difference may be more significant and thus suitable for node splitting (Yu et al., 2011). More specifically, the optimisation is to locate a feature pair that produces the largest variance on their difference. The split threshold is the mean of feature difference. For evaluating the goodness of each candidate split, the main computational cost is on variance estimation, which has a complexity of $Q(n)$, with n the data sample number. This is more expensive than classification forests (Equation (2.3)) and pseudo sample based strategy (Breiman, 2001), while similar to regression forests (Equation (2.6)). A similar but still different split method is to evenly divide the input training data by random projection (Perbet et al., 2009). Because the objective is to separate data evenly in each tree node, the split threshold is set as the median of projected values, without the need for any parameter optimisation.

The third is the unsupervised entropy for constructing density forest (Criminisi and Shotton, 2012). This criterion assumes Gaussian distributed data and computes information gain with the determinant of covariance matrix, which is related to the volume of the hyperellipsoid that bounds the uncertainty of data distribution (Sim and Roy, 2005). Given n d-dimensional data samples in a node, the complexity of computing the covariance matrix is $O(nd^2)$. Therefore, this may be computationally expensive when the feature dimensionality and sample size are large. Additionally, while working well on low-dimensional data with a full rank covariance matrix, it may suffer from the rank deficiency problem particularly in high-dimensional cases. That is, this scheme fails when the rank of covariance matrix is lower than the feature dimensionality or sample number, e.g. zero-valued determinant.

To address this rank-deficiency problem, Pei et al. (2013) consider a split criterion based on the trace measure of covariance matrix and a scatter index. In practice, the former's complexity is $Q(nd)$ because there is no need to compute the whole covariance matrix rather than only the variance of each dimension. The computational complexity of the scatter index is $O(nd)$ as well. Therefore, this split optimisation strategy is more efficient than that (i.e. $O(nd^2)$) of (Criminisi

and Shotton, 2012), particularly so on high-dimensional data. For optimising a split node, they select the feature pair and threshold corresponding to the largest information gain computed with their introduced criterion, similar to the split function in (Yu et al., 2011).

2.2.4 Weak Learners – Split Functions

The split function is one of the most crucial components in random forests. Their design largely depends on the specific target problems and application settings. Many different types of split functions have been developed for solving a variety of computer vision applications such as image categorisation, object detection, pose estimation, face analysis and so forth. Here, we review and discuss notable works presented recently. A common split function is axis-aligned (or axis-parallel) linear function (Breiman, 2001; Criminisi and Shotton, 2012), e.g. Equation (2.1). Specifically, a threshold (ϑ_2) is selected on a specific feature variable (x_{ϑ_1}), which can seen as a split line. Samples are routed to the left or right child node according to whether being lower than the threshold. This weak learner is also referred as stump (Viola and Jones, 2004). One generalisation is the combination of multiple split lines. In the proposed forest models, we choose this simple weak learner (i.e. single-feature split) because it is simple and efficient in execution, also generalises to more complex cases. More generally, oblique linear splits are another split model (Heath et al., 1993; Ho, 1995; Menze et al., 2011). The defined split decision hyperplanes are not necessarily aligned with any axis of the feature space. The execution speed depends on the complexity of hyperplanes.

Alternatively, stronger weak models can be obtained by using non-linear split functions, such as via margin-maximisation (Ho, 1998; Yao et al., 2011), boosting (Yin et al., 2007). Many linear and non-linear split functions are evaluated on the original input features. However, other features can be synthesised and utilised, e.g. PCA-based features and variance (Fanello et al., 2014), feature pairwise difference (Yu et al., 2011; Pei et al., 2013), In some forest variants, multiple feature types many be considered, and at a time only one type is randomly selected in each node (Fanello et al., 2014). Besides, multiple optimisation objectives are possible and required to realise in a single forest model (Gall et al., 2011; Yang and Patras, 2013; Tang et al., 2013; Doumanoglou et al., 2014; Schulter et al., 2014).

Recently, a variety of split function learning algorithms have been developed in diverse contexts. They include, learning in an entangled setting where intermediate classifier output is stacked with the the original input data (Montillo et al., 2011; Kontschieder et al., 2013); glob-

ally optimising (Schulter et al., 2013a,b; Kontschieder et al., 2015); training in a semi-supervised manner (Tang et al., 2013; Leistner et al., 2009); optimising with the SVM learner (Yao et al., 2011; Marin et al., 2013); learning in a stage-wised and coarse-to-fine way (Tang et al., 2014); hierarchically optimising in multi-task settings (Zhao et al., 2014c; Doumanoglou et al., 2014); conditionally learning by modelling the dependency between the target variables and a global latent variable (Dantone et al., 2012; Sun et al., 2012); learning based on nearest class mean classifier (Ristin et al., 2014, 2015); optimising using randomised multi-layer perceptrons (Rota Bulo and Kontschieder, 2014); training by back propagation (Kontschieder et al., 2015). For more details, I refer the reader to the corresponding papers.

Next, let us return to the main problems and topics of this thesis, after reviewing the background techniques on random forests. As a common data form in our physical world, images and videos considered in various problems can be observed and collected by either a single or multiple camera views, e.g. in usual video surveillance scenarios. In either case, visual content analysis can be performed with one of the three classic learning strategies, depending on the availability of data labels. For simplicity, this chapter divides the literature remainder into two parts from data source perspective: single-camera visual data structure discovery (Section 2.3) and multi-camera visual data structure discovery (Section 2.4), with the aim to provide a broad foundation and context for the studies presented in this thesis.

2.3 Single-Camera Visual Data Structure Discovery

Most studies on visual data structure analysis are devoted to the single camera view setting. This section discusses a number of seminal learning methods and techniques for single-camera data cluster structure analysis and discovery. In particular, this section is separated into three subsections based on the data form and label availability:

1. Unsupervised visual data structure discovery (e.g. clustering): Particularly, each visual data sample/pattern is described by some descriptor such as feature vectors and no data annotation is given, as shown in Figure 2.1-(d) (Section 2.3.1).
2. Semi-supervised visual data structure discovery (e.g. constrained clustering): Partial/incomplete sparse annotations additional to data features are accessible. Instead of individual class labels as in classification, pairwise constraints (Figure 2.1-(c)) over data samples are commonly offered in clustering (Section 2.3.2).

3. Multi-source data structure discovery (e.g. multi-way clustering): Beyond visual feature data, auxiliary data from other non-visual sources are also available (Section 2.3.3).

2.3.1 Unsupervised Visual Data Structure Discovery

Organising visual data into coherent and meaningful cluster structures without any supervision is one of the most fundamental strategies in data analysis. This is also known as *clustering* or *cluster analysis*. In this learning setting, one may have no access to annotations or supervisions that tag data samples with some identifiers, e.g. category labels or numerical measures. The absence of annotation information distinguishes visual data cluster analysis (unsupervised learning) from discriminant analysis (supervised learning) such as classification, and regression. The goal of cluster analysis is to find the structure in visual data, more precisely, to discover whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics (Online-Dictionary, 2015). Clustering is thus exploratory in nature. So far, it is still challenging to design a general purpose clustering algorithm (Jain, 2010). Whilst human is an excellent cluster seeker in two or three dimensions, automated clustering algorithms are needed for higher-dimension cases, e.g. visual data. Particularly, mining visual data structure, e.g. images or videos, is often more difficult due to the inherent nature of high dimensionality and inevitable noisy/inaccurate feature representations. These challenges have been constantly driving the research effort on cluster analysis algorithms in computer vision, pattern recognition fields.

In general, the main purposes of data cluster structure analysis are (Jain, 2010):

1. *Underlying structure*: To gain insights into visual data and detect outliers or inliers.
2. *Natural classification*: To measure and identify the proximity relationships among data samples.
3. *Compression*: To organise visual data in form of clusters and summarise it, e.g. video summarisation in visual surveillance presented in Chapter 5.

Beyond a number of important generic clustering algorithms reviewed and summarised below, more extensive studies on data clustering techniques and methods can be found in (Jain and Dubes, 1988; Duda et al., 2012; Hartigan, 1975; Sokal et al., 1963; Han et al., 2011). Clustering methods can be broadly grouped into two categories: (1) *Hierarchical* clustering algorithms: finding nested clusters either in agglomerative mode (forming a cluster hierarchy by starting with

each data sample in its own cluster and gradually merging the most similar pair of clusters) or in divisive mode (to begin with all data samples in one cluster and recursively splitting each cluster into smaller ones). The input is a pairwise similarity matrix. Well-known algorithms include single-link or nearest neighbour (McQuitty, 1957; Sneath, 1957; Gower and Ross, 1969; Sibson, 1973), complete-link (Defays, 1977; Hansen and Delattre, 1978). (2) *Partitional* clustering algorithms: finding all the clusters simultaneously as a data partition without a hierarchical structure. The input includes either a similarity matrix or a data matrix. The most famous algorithm is k-means (Steinhaus, 1956; Lloyd, 1982; Ball and HALL DJ, 1965; MacQueen et al., 1967). Among the clustering literature, a number of notable clustering algorithms are dense based methods (Frank and Todeschini, 1994; McLachlan and Basford, 1988; Ester et al., 1996; Blei et al., 2003; Welling et al., 2004), subspace clustering algorithms (Agrawal et al., 1998), graph theoretic or spectral clustering (Hagen and Kahng, 1992; Shi and Malik, 2000; Meila and Shi, 2001; Ng et al., 2002; Belkin and Niyogi, 2001; Pavan and Pelillo, 2007), and information theoretic formulation based methods (Roberts et al., 2001; Tishby et al., 2000).

In spite of grade stride made in the last fifty years since k-means, data clustering remains a difficult problem, particularly for complex data. One fundamental challenge lies in defining an appropriate similarity measure (Jain, 2010). The follows are converged to studies with regards to the data similarity measure issue for achieving robust cluster structure discovery.

Approaches to adapting local data structures (or local neighbourhoods) for improving the accuracy and robustness of similarity or affinity matrices have been presented in (Zelnik-manor and Perona, 2004; Wang et al., 2008). Particularly, their focus has been spent on learning an adaptive scaling factor σ for the Gaussian kernel (also known as radial basis function or heat kernel) $\exp\left(-\frac{\text{dist}^2(\mathbf{x}_i,\mathbf{x}_j)}{\sigma^2}\right)$, when computing the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$, with $\text{dist}(\cdot,\cdot)$ pairwise distance and σ the brand width parameter. These methods, however, are still susceptible to the presence of noisy and irrelevant features.

To mitigate the above issue, Pavan and Pelillo (2007) proposed a graph-theoretic algorithm for forming tight neighbourhoods via selecting the maximal cliques (or maximising average pairwise affinity), with the hope of constructing graphs with fewer false affinity edges between samples. More recently, a k nearest neighbour (k-NN) based similarity graph generation method is developed by Premachandran and Kakarala (2013). Specifically, the consensus information cumulated from multiple k-NNs is utilised in this algorithm for discarding noisy edges and iden-

tifying strong local neighbourhoods.

In contrast to all the aforementioned methods that blindly trust all available variables, the proposed graph inference method presented in Chapter 3 exploits discriminative and informative features for attaining more robust data pairwise similarities. The resulting affinity matrix is thus more robust against the inherent noise in real-world visual data. Random forest-based affinity graph construction has been attempted in (Shi and Horvath, 2006; Criminisi and Shotton, 2012; Zhu et al., 2013). The intuition is that tree leaf nodes contain discriminative data partitions, which could be exploited for generating robust affinity graphs. It is showed that the above approaches are special cases of the proposed affinity inference method. Specifically, a generalised model is derived, which is not only capable of learning discriminative feature subspaces for robust affinity graph construction as in previous methods, but also able to further exploit the hierarchical structure of random forest to better capture subtle and weak data proximity.

2.3.2 Semi-Supervised Visual Data Structure Discovery

Unsupervised visual data structure discovery or data cluster analysis is inherently ill-posed, i.e. the similarity definition is not explicitly specified. Therefore, it is generally hard for many clustering algorithms to generate desired clusters. In cases, additional side information is available from human experts. An appropriate utilisation of the external information along with data features helps in finding satisfactory data cluster memberships. One common side information is expressed in form of pairwise constraint that specifies a pair of samples should be assigned with the same cluster (must-link) or with two different clusters (cannot-link). These prior knowledge is a type of weak supervision compared to category labels since they are not sufficient to infer the explicit classes. Clustering with such external information is also known as *constrained clustering* (Wagstaff et al., 2001; Basu et al., 2004a) or *semi-supervised clustering* (Basu et al., 2002; Kulis et al., 2009; Araujo, 2015). This is because that this learning setting is similar in spirit to semi-supervised learning (Blum and Mitchell, 1998; Chapelle et al., 2006; Zhu et al., 2003) where only a limited number of data rather than the whole dataset have supervision information, i.e. some sample pairs are constrained with either must-link or cannot-link whilst the (most) others are not. Constrained clustering aims to exploit this prior belief as constraints (or weak supervision) over data feature information to influence the clustering process so as to obtain a data structure more closely resembling human perception or desired high-level explanation. Below, recent classic constrained clustering methods will be briefly reviewed and discussed. For

a complete survey on semi-supervised learning beyond constrained clustering, Zhu (2005) and Chapelle et al. (2006) provide a more general and systematic discussion.

There are generally two paradigms to exploit pairwise constraints for semi-supervised data structure discovery. The first paradigm is distance metric learning (Xing et al., 2002; Yang and Jin, 2006; Weinberger and Saul, 2009; Der and Saul, 2012; Ying and Li, 2012), which learns a distance metric that respects the constraints, and runs ordinary clustering algorithms, such as k-means, with distortion defined in the learned metric. The second paradigm adapts directly existing clustering methods, such as k-means (Wagstaff et al., 2001; Basu et al., 2008) and spectral clustering methods (Wang et al., 2012c,b) to satisfy the given pairwise constraints. Instead of looking from the above strategy angle this section discusses the related semi-supervised clustering methods from two practical challenges: (1) sparse constraints; and (2) imperfect annotators/oracles.

Sparse Constraint Propagation Approaches that perform similarity matrix based constrained clustering generally follow a procedure that first manipulates pairwise data affinity/similarity with constraints and then applies existing clustering algorithm, e.g. spectral clustering. For instance, Kamvar et al. (2003) trivially adjust the elements in an affinity matrix with "1" and "0" to respect must-link and cannot-link constraints, respectively. No constraint propagation is considered in this method, e.g. measuring the similarity between unconstrained data pairs is not benefited from the given constraints. It is thus not effective particularly in case of sparse constraints.

The problem of sparse constraint propagation has been considered in previous studies. Specifically, Lu and Carreira-Perpinán (2008) proposed to perform propagation with a Gaussian process. This method is limited to the two-class problem, although a heuristic approach for multi-class problems is also discussed. Li et al. (2009) formulated the propagation problem as a semi-definite programming (SDP) optimisation problem. The method is not limited to the two-class problem, but solving the SDP problem involves extremely large computational cost. In (Yu and Shi, 2004), the constraint propagation is also formulated as a constrained optimisation problem, but only must-link constraints can be employed. In contrast to the above methods, the proposed approach described in Chapter 4 is capable of performing effective constrained clustering using both available must-links and cannot-links, whilst it is not limited to two-class problems.

The state-of-the-art clustering results are achieved by Lu and Ip (2010); Lu and Peng (2013a).

They address the propagation problem through manifold diffusion by decomposing the constraint propagation problem into two independent semi-supervised learning sub-problems (Zhou et al., 2004; Zhu et al., 2003). The locality-preserving character in learning a manifold with dominant eigenvectors makes the solution less susceptible to noise to a certain extent, but the manifold construction still considers the full feature space, which may be corrupted by noisy features. Chapter 4 shows that the manifold-based method is not as effective as the proposed model featured with discriminative-feature driven constraint propagation. Moreover, the methods (Lu and Ip, 2010; Lu and Peng, 2013a) as well as other methods ((Yu and Shi, 2004; Lu and Carreira-Perpinán, 2008; Li et al., 2009)), do not have a mechanism to handle noisy constraints.

All these constrained clustering methods above mostly considers single modality data. On the other hand, we have access to an increasing number of multi-modal visual data, mainly due to the proliferation of social media websites, e.g. YouTube, Facebook, and Flickr. Specifically, images and videos extracted from such website sources are often associated with related text and tag labels which provide rich and semantically meaningful perspectives complementary to visual content. Effectively modelling visual and text modalities jointly may bring into additional benefits, as shown in (Lu and Peng, 2013b, 2012; Fu et al., 2011, 2012; Yang et al., 2014). More specifically, Fu et al. (2011) proposed a unified multi-modal constraint propagation method with a closed-form solution, where an individual graph is built for each data modality (e.r. visual or textual features) and a random walk is defined across graphs. On such heterogeneous graphs, a random walk process is defined and multiview label propagation (Zhou and Burges, 2007) is then applied to solve decomposed subtasks. Lu and Peng (2012) consider jointly both homogeneous and heterogeneous constraint propagation over multiple modalities by a new constrained sparse representation method in the context of cross-modal retrieval. Similarly, Lu and Peng (2013b) present a unified framework for intra-view and inter-view constraint propagation by decomposing the two types of constraints into semi-supervised learning sub-problems. Intra-view constraints are utilised to refine and improve intra-view similarity measures, which in turn benefit the subsequent inter-view constraint propagation. Further, Yang et al. (2014) propose a low rank based matrix completion algorithm for cross-view constraint diffusion for better preserving both local and global data structures. Zhang et al. (2015) design a multi-view constrained clustering algorithm in the framework of non-negative matrix factorisation. However, all these methods assume the correctness of all given pairwise constraints, either intra-view or cross-view.

Handling Imperfect Oracles A small number of constrained clustering studies consider imperfect oracles or noisy constraints whereas most simply assume perfect constraints available. The usefulness of constraints in clustering performance is discussed and investigated in (Wagstaff et al., 2006; Davidson et al., 2006). In particular, two constraint set properties are proposed: (1) "Informativeness", which refers to the amount of information a constraint can provide for some specific clustering algorithm, thus algorithm dependent; (2) "Coherence", which measures the agreement degree (e.g. projected overlap) between must-link and cannot-link under a distance metric, so only partial constraint pairs can be exploited. Nevertheless, both measures are on the whole constraint set rather than on individual constraints, so providing no information about what are noisy and ill pairwise links. Moreover, no concrete method is proposed to exploit such metrics for improved constrained clustering, except offering some indication about the utility of constraint sets on the future clustering performance. Zeng et al. (2007) and Ares et al. (2012) demonstrate the negative effectiveness of incorrect constraints on clustering accuracy. Whereas Freund et al. (2008) analyse analytically and empirically the effect of noisy constraints on constrained clustering algorithms in the random graph theory framework. More recently, Van Craenendonck and Blockeel (2015) investigate the informativeness and coherence metrics proposed in (Davidson et al., 2006) for explaining the clustering performance difference among a number of methods and find some limitation w.r.t. these two measures and selection clustering methods given a dataset.

A few specific approaches to handling noisy constraints have also been proposed. Nelson and Cohen (2007) extend the chunklet model (Strehl et al., 2000) to soft constraints for acquiring additional robustness against constraint errors. The key idea is to punish constraint violation by constraint sampling in order to avoid the pitfalls of local approximation. This method implicitly assumes a large number of pairwise constraints in noisy link detection. Also, human confidence is assumed related to the correctness of constraints, which is used in constraint violation penalty weighting. Coleman et al. (2008) proposed a constrained spectral clustering algorithm capable to deal with inconsistent constraints. The main idea is to minimise the number of must-links between clusters and the number of cannot-links within clusters. Specifically, pairwise links are transformed into subspaces which are set as the allowed space for spectral clustering solutions. This model is restricted to only the two-class problem setting due to the adoption of two-correlation clustering idea. Similar to (Nelson and Cohen, 2007), a large number of pairwise

constraints are required in inconsistency detection since explicit and direct interaction between constraints are needed via shared constrained samples.

Beyond constrained clustering, the problem of imperfect oracles has been explored in active learning (Donmez and Carbonell, 2008; Du and Ling, 2010; Yan et al., 2011; Sogawa et al., 2013) and online crowd-sourcing (Kittur et al., 2008; Welinder and Perona, 2010). This thesis (in Chapter 4) presents a method that differs significantly from these studies as what concerned in this method is identifying noisy or inconsistent pairwise constraints rather than inaccurate class labels.

2.3.3 Multi-Source Data Structure Discovery

The data considered in the above two data structure discovery scenarios is associated with only one source, e.g. the visual source. In addition, each single data sample may be associated with multiple information sources, e.g. apart from visual features, an individual video is also linked with additional textual descriptors that are drawn from other correlated and independent sources (Figure 1.5), potentially possible in video surveillance. Different source data can be significantly distinct in representation and statistical distributions and heterogeneous to one another. This is also called *multi-way clustering* (Jain, 2010). Whilst a straightforward way is to pull all source features into a combined vector ahead of clustering, this representation is neither natural nor coherent and may result in poor clusters. It is desirable to derive appropriate joint learning algorithms for the heterogeneous source data clustering setting. According to data source, one can further group this multi-way setting into two sub-categories:

1. *Multi-modality data learning*: where different types (or modalities) of feature data are computed from the same single source for capturing different perspectives of the same information, e.g. visual colour and texture.
2. *Multi-source data learning*: where multiple different physical sources are involved, each may encode some particular type of information.

For the application of multi-way clustering, video summarisation or compression is selected in this work (Chapter 5) due to: (1) Video summarisation is a fundamental visual surveillance task; (2) Data compression is one main purpose of cluster analysis. Below, this section briefly reviews existing multi-modality and multi-source data learning approaches (mostly in the computer vision domain), and finally video summarisation methods.

Multi-Modality Data Learning There exist studies that exploit different sensory or information modalities from a single source for data structure mining. Many earlier studies (Barnard and Forsyth, 2001; Blei and Jordan, 2003; Blei et al., 2003; Duygulu et al., 2002; Lavrenko et al., 2003) mainly focus on co-occurrence relationship learning between visual parts (e.g. image regions) and text. This type of approaches assume (1) the availability of part level data annotation during the training stage, and (2) the existence of correct associations. Both requirements however can be largely invalid in real-world applications.

An alternative is to learn a joint latent space wherein paired visual samples and text are projected to nearby locations. In this way, nearest neighbour methods can be used to reason uncertainty and infer semantics. Many of these embedding based approaches rely on the Canonical Correlation Analysis (CCA) algorithm (Hotelling, 1936). Hardoon et al. (2004) and Rasiwasia et al. (2010) applied CCA to learn the latent shared space for images and text. Blaschko and Lampert (2008) developed a cross-modal spectral clustering algorithm based on Kernelised CCA (KCCA). Udupa and Khapra (2010) and Vinokourov et al. (2002) utilised CCA for cross-language retrieval.

The third strategy is multi-view embedding techniques, in which visual samples are characterised by visual and text views. Multi-view learning methods include generalisations (Yakhnenko and Honavar, 2009; Rai and Daume, 2009; Sharma et al., 2012) and extensions (Gong et al., 2014b; Fu et al., 2015) of CCA/KCCA, multi-view metric learning (Quadrianto and Lampert, 2011), large margin predictive latent subspace learning (Chen et al., 2012), unsupervised deep learning methods such as Restricted Boltzmann Machine (RBM) (Srivastava and Salakhutdinov, 2012) and auto-encoders (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012). For data clustering, Cai et al. (2011) proposed to perform multi-modal image clustering by learning a commonly shared graph-Laplacian matrix from different visual feature modalities. Heer and Chi (2001) combined linearly individual similarity matrices derived from multi-modal webpages for web user grouping. Karydis et al. (2009) presented a tensor based model to cluster music items with additional tags. In terms of video analysis, the auditory channel and/or transcripts have been widely explored for detecting high-level concepts from multimedia videos (Zhang et al., 2004; Fu et al., 2014), summarising highlights in news and broadcast programs (Taskiran et al., 2006; Gong, 2003), or locating speakers (Khalidov et al., 2011). User tags associated with web videos (from moment-sharing and social multimedia websites like YouTube, Flickr and Facebook) have

also been utilised in (Wang et al., 2010; Toderici et al., 2010; Wang et al., 2012a).

In contrast to all these methods above, surveillance videos captured from public spaces are typically without auditory signals nor any synchronised transcripts and user tags available. Instead, this thesis explores alternative non-visual data drawn independently elsewhere from multiple sources, with inherent challenges of being inaccurate and incomplete, unsynchronised to and may also be in conflict with the observed visual data (Chapter 5).

Multi-Source Data Learning One possible multi-source data structure mining solution can be clustering ensemble (Strehl and Ghosh, 2003; Topchy et al., 2005; Jain, 2010) where a collection of clustering instances is generated and then aggregated into the final clustering solution. Typically only single data source is considered, but it can be easily extended to handle multi-source data, e.g. creating a respective clustering instance for each source. Nonetheless, cross-source correlation is ignored since the clustering instances are separately formed and no interaction between them is involved. A closer and competitive approach to our model presented in Chapter 5 is the Affinity Aggregation Spectral Clustering (AASC) (Huang et al., 2012), which learns data structure from multiple types of homogeneous information (visual features only). Their method generates independently multiple affinity data matrices by exhaustive pairwise distance computation for every pair of samples in every data source. It suffers from unwieldy representation given high-dimensional data inputs. Importantly, despite that it seeks for optimal weighted combination of distinct affinity matrices, it does not consider correlation between different sources in model learning, similar to clustering ensemble (Strehl and Ghosh, 2003; Topchy et al., 2005). Differing from the above models, the proposed Multi-Source Clustering Forest (see Chapter 5 for details) overcomes these problems by generating a unified single affinity matrix that captures latent correlations among heterogeneous types of data sources. Furthermore, our forest model has a unique advantage in handling missing non-visual data over (Strehl and Ghosh, 2003; Topchy et al., 2005; Huang et al., 2012).

Video Summarisation Automated video summarisation facilitates a holistic understanding of long videos in a short time by generating a compact summary composed of important/key content, particularly with surveillance videos captured by cameras that operate all the time (Xiong et al., 2006; Truong and Venkatesh, 2007). One common way to summarise redundancy videos is to identify and combine key frames, shots or objects. The discovery of these key contents requires the underlying data structure analysis, e.g. importance prediction, cluster discovery, for

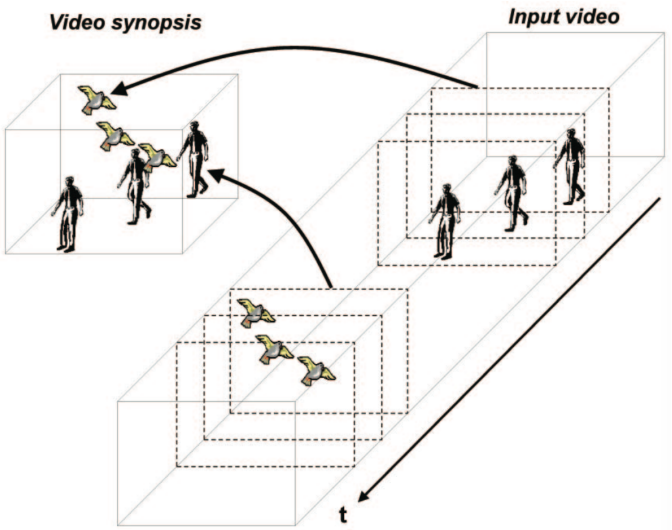


Figure 2.4: An example for illustrating the summarisation process based on object trajectories distributed over different times in video streams. This is borrowed from (Pritch et al., 2008).

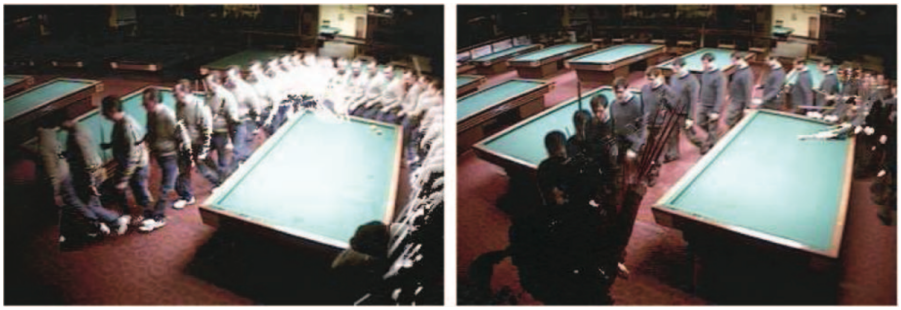


Figure 2.5: Examples of detected object trajectories in a "Billiard" video (Pritch et al., 2008).

modelling the content redundancy patterns (Truong and Venkatesh, 2007). Below, classic video summarisation methods are analysed. For a general review on this problem, Xiong et al. (2006) and Truong and Venkatesh (2007) provide a comprehensive coverage on common techniques and models.

Contemporary video summarisation methods can be broadly classified into two paradigms: key-frame-based (Lee et al., 2012; Wolf, 1996; Zhang et al., 1997; Truong and Venkatesh, 2007; Money and Agius, 2008) and object-based (Kang et al., 2006; Rav-Acha et al., 2006; Pritch et al., 2007, 2008, 2009; Wang et al., 2011; Feng et al., 2012) methods. The key-frame-based approaches aim to select representative key-frames for building a storyboard of still images as video summary. This is typically achieved by analysing low-level imagery properties, e.g. optical flow (Wolf, 1996) or global scene colour difference (Zhang et al., 1997). Alternatively, Lai and Yi (2012) and Ma et al. (2005) proposed to seek for key-frames by using human attention models.

Recently, Lee and Grauman (2015) presented an approach to estimating the importance scores of image regions such as objects or people, which are then utilised as key-frame selection criteria for summarising egocentric videos. Whilst the key-frame based summary is a straightforward video summarisation mechanism, it is inherently limited and less rich in representation due to the loss of dynamic information encoded in original videos.

Object-based techniques (Rav-Acha et al., 2006; Pritch et al., 2007, 2008; Feng et al., 2012), on the other hand, rely on object segmentation and tracking to extract object-centric trajectories/tubes, and compress those tubes to reduce spatio-temporal redundancy. Therefore, these summary videos can retain much dynamic motion information of the raw videos when compared with the former paradigm. Specifically, Kang et al. (2006) exploited a spatial-temporal contrast based video saliency detection method to extract informative space-time visual regions. The final montage video summary is established by a space-time region merging algorithm which allows both spacial and temporal shifting for maximising video compression ratio. Similar summarisation principles are adopted by (Rav-Acha et al., 2006; Pritch et al., 2007, 2008; Feng et al., 2012) with some newly introduced features and capabilities, e.g. (Pritch et al., 2007, 2008) allows to summarise endless video streams by generating video summary/synopsis of user-specified time duration over a particular video stream range, whilst the algorithm in (Rav-Acha et al., 2006) only enables to summarise short videos. A summarisation illustration example and extracted object trajectories are shown in Figure 2.4 and Figure 2.5 respectively. On the other hand, the method presented in (Feng et al., 2012) is characterised with online real-time chronological video summarisation, e.g. rapid object trajectory extraction, low memory requirement, and preserved temporal order of objects.

Both the above schemes utilise solely visual information and make implicit assumptions about the completeness and accuracy of the visual data available in extracting features or object-centred representations. They are unsuitable and unscalable to complex scenes where visual data are inherently incomplete and inaccurate, mostly the case in surveillance videos. The proposed method in Chapter 5 differs significantly to these studies in that it exploits not only visual data without object tracking, but also non-visual sources as complementary information. The summary generated by the proposed approach is semantically enriched – it is labelled automatically with tags, e.g. traffic condition, weather, or event. All these tags are learned from heterogeneous non-visual sources in an unsupervised manner during model training without any manual labels.

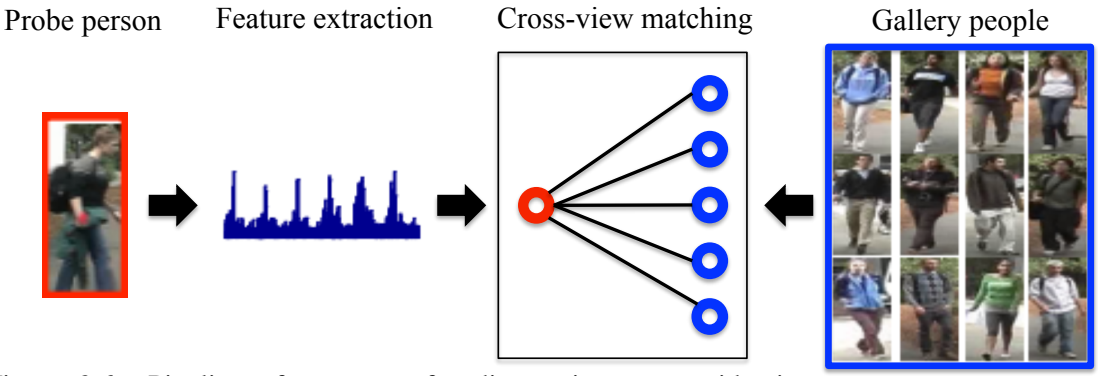


Figure 2.6: Pipeline of a system for discovering person identity structure or person re-identification.

2.4 Cross-Camera Visual Data Structure Discovery

The previous section discusses visual data structure discovery and cluster analysis for single-camera settings. Whilst most existing approaches to data analysis and correlation modelling are devoted to single camera view settings (Rodriguez et al., 2011; Amer and Todorovic, 2011; Ryoo, 2011; Gaur et al., 2011), extending these methods to scenarios with multiple disjoint cameras is non-trivial due to the unknown inter-camera time gaps and significant appearance variations across camera views. This section mainly focuses on reviewing existing approaches to person identity structure discovery across non-overlapping camera views, a fundamental multi-camera surveillance problem.

2.4.1 Person Identity Structure Discovery

For making sense of the vast quantity of video data generated by large scale surveillance camera networks in public spaces, automatically associating and recognising individual persons across non-overlapping camera views distributed at different physical locations is essential. This task is also known as *person re-identification* (ReID). This ability enables automated discovery and analysis of person-specific long-term structural activities over widely expanded areas and is fundamental to many other important surveillance applications such as multi-camera people tracking and forensic search. Specifically, person re-identification aims to match people across non-overlapping camera views over different space and time (Gong et al., 2014a). Typically, person ReID is performed by matching cross-view single or multiple images, called *single-shot* person re-identification (Section 2.4.1) and *multi-shot* person re-identification (Section 2.4.1). A number of important person ReID studies are discussed in this section and more complete reviews can be found in (Gong et al., 2014a; Vezzani et al., 2013; Bedagkar-Gala and Shah, 2014).

Single-Shot Person Re-Identification

Generally, a person ReID system (Figure 2.6) involves two key components: (1) feature representations of people; and (2) matching model.

Feature Representations Designing a suitable person ReID feature representation is essential and challenging. Ideally, the feature should be identity discriminative under even large cross-view changes in illumination, view point, human pose, background clutter and occlusion. A variety of ReID features have been proposed recently, including colour, texture, gradient, edge, shape, global or localised features. Most ReID methods generally exploit multiple such appearance features due to their complementary effects and the lack of uniformly effective feature types (Gray and Tao, 2008; Farenzena et al., 2010; Hirzer et al., 2012; Zheng et al., 2013; Liu et al., 2014b; Paisitkriangkrai et al., 2015). The common representation form is bag-of-words, which can be easily concatenated for achieving a combination of multiple features.

Spatial structure information of people's appearance is an important cue. Many different spatial decomposition schemes have been developed to integrate the spatial configuration into the feature representation, e.g. triangulated graphs (Gheissari et al., 2006), uniform horizontal strips (Gray and Tao, 2008; Layne et al., 2012; Prosser et al., 2010; Zheng et al., 2013; Liu et al., 2012), concentric rings (Zheng et al., 2009), or localised patches (Zhao et al., 2013b,a, 2014a; Liu et al., 2014b; Li et al., 2014a; Bak et al., 2010; Zheng et al., 2015; Paisitkriangkrai et al., 2015; Liao et al., 2015). In case that the body topology structure (part configuration) can be detected using human parsing and pose estimation techniques, more robust and relevant features from detected body parts can be computed and matched, simultaneously alleviating the negative contamination by background. For example, the feature designing method presented in (Farenzena et al., 2010) utilises the principles of symmetry and asymmetry in human body structure to segregate the perceptually meaningful body parts as foreground from the whole image. Specifically, higher importance weights are imposed to regions around the vertical symmetry axis than those far away from it. This allows the suppression of distractive background information during the feature matching procedure. Additionally, a part-based (e.g. head, torso, arms and legs) ReID method (Cheng et al., 2011) allows selective matching between pairs of localised body part rather the entire person. This scheme not only partially filters background noise, but also offers robustness to partial and self occlusion.

In addition to hand-crafted low-level visual features, saliency information (Liu et al., 2012;

Zhao et al., 2013b,a; Wang et al., 2014a; Zhao et al., 2014b) is recently extracted and exploited for person ReID. The saliency weights about appearance are learned from some unlabelled reference images for capturing the localised protuberance statistics among a population. Such knowledge shows exceptional robustness against large cross-camera viewing condition variations.

In contrast to machine vision often using continuous and high-dimensional features, human may perform person ReID using discrete and low-dimensional appearance attributes, a more abstract and robust way. Discrete attributes can be more reliable and unambiguous in inference, such as shoe-type, hair-style, clothing-style (Lampert et al., 2009). Such attributes can be less variable against the photometric and geometric transformations across camera views, compared with continuous appearance futures. On the other hand, this mid-level 'semantic attribute representation' is fairly similar to the descriptor communicated verbally among people for specifying people instances in reality. Attribute representations can be computed and abstracted from the widely-used low-level features, e.g. support vector machine trained with additional attribute annotations (Layne et al., 2012, 2014a; Li et al., 2014a). Whilst the attribute ontology is often manually specified by human experts, it can be also automatically mined from large scale internet data (Layne et al., 2014b). Importantly, this method is characterised by annotation free, richer diversity and better generalisation as demonstrated in (Layne et al., 2014b). Semantic attribute representation possesses a number of benefits: (1) Being more powerful than raw features as the pre-trained attribute classifiers learn the variances in appearance of each attribute as well as invariance to appearance of the corresponding attribute across camera views (Lampert et al., 2009; Siddiquie et al., 2011; Liu et al., 2011). (2) Complement low-level features for building more powerful representation (Layne et al., 2012; Liu et al., 2011). (3) Being suitable for direct human-robot interaction, e.g. allowing people search using human-attribute-profiles (Kumar et al., 2011; Layne et al., 2014a).

Recently, deep learning visual features for ReID has also been attempted (Li et al., 2014b; Yi et al., 2014; Ding et al., 2015), inspired by its massive success in a wide range of computer vision tasks. Unlike hand-crafted features, these neural network based methods learn person discriminative features from raw visual data. Specifically, a filter pairing neural network is particularly designed for jointly handling the misalignment, photometric and geometric transforms, occlusions and background clutter issues in ReID (Li et al., 2014b). Yi et al. (2014) applied a symmetric 'siamese' neural network to learn ReID features that are robust to the inherent challenging

cross-view variances. More recently, a triplet-based ReID feature learning model is derived in (Ding et al., 2015), where each triplet unit contains a query image from one view, a true and false match from another view. The objective is to learn a convolutional network that maximises the relevance distance between the matched pair and the unmatched pair. One main weakness of deep learning based ReID methods is the requirement of many training data for avoiding the over-fitting problem (Krizhevsky et al., 2012; Li et al., 2014b).

Model Learning When pairwise labelled data are available, one can learn an identify sensitive appearance transfer function or distance metric for modelling the cross-view photometric and geometric transformations. An intuitive method is to learn the Brightness Transfer Functions (BTF) between two camera views, with the aim to capture the changes in the colour distributions of object travelling from one view to another (Porikli, 2003; Prosser et al., 2010; Chen et al., 2008; D'Orazio et al., 2009; Javed et al., 2008; Jeong and Jaynes, 2008; Lian et al., 2012). These methods typically assume the availability of perfect foreground detections, which however is largely invalid in practical cases. Furthermore, the actual transfer functions between views may be complex and multi-modal, which is very difficult to be approximated by a single BTF function, due to many inherent variation factors such as pose, background, lighting. Li and Wang (2013a) proposed a multi-modal model learning scheme to alleviate this problem.

A more popular alternative is distance metric learning. The main idea of metric learning is to optimise the model parameters so that the cross-view inter-person distance is large whilst intra-person distance is small, i.e. person identity discriminative. Existing metric learning methods include Large Margin Nearest Neighbour (Weinberger et al., 2005), Information Theoretic Metric Learning (Davis et al., 2007), Logistic Discriminant Metric Learning (Guillaumin et al., 2009), KISSME (Kostinger et al., 2012), RankSVM (Prosser et al., 2010), Probabilistic Relevance Distance Comparison (Zheng et al., 2013). Most of these are Mahalanobis metric learning, which need the optimisation of a full matrix. Whilst the RankSVM model constrains only a single weight parameter for each feature dimension, which is thus potentially less effective than the Mahalanobis distance.

Often, supervised model learning requires a large number of exhaustively labelled data. This assumption significantly limits their scalability in real-world scenarios, e.g. the need of collecting sufficient pairwise labels for each of many camera pairs. One solution is semi-supervised learning (Loy et al., 2013; Liu et al., 2014b, 2013; Figueira et al., 2013) that can exploit the

manifold geometry structures of unlabelled data for constraining the learning process given only very sparse labels. An alternative perspective to solve the label data scarcity problem is transfer learning and domain adaptation. Particularly, one wishes to learn a ReID model for a target camera pair with only a small number of labelled people or even no annotation. To this end, a model pre-trained from other auxiliary data is exploited and / or adapted to the new target dataset. Recently, Hu et al. (2015) developed a deep transfer metric learning algorithm in the neural network framework to transfer the discriminative knowledge from labelled auxiliary datasets to the unlabelled target dataset. This model is trained by enforcing two constraints as: (1) maximising the inter-class/person variations, and minimising the intra-class variations; (2) minimising the divergence between source and target domains. Adapting and transferring ReID model is a challenging problem and remains open although some initial efforts have been made (Layne et al., 2013; Wu et al., 2013; Ma et al., 2013; Shi et al., 2015; Hu et al., 2015).

Multi-Shot Person Re-Identification

In many cases, more than one person shot can be accessible. Multiple images of the same person have been exploited for person re-identification. For example, Gheissari et al. (2006) generated a decomposable spatio-temporal graph for identifying localised regions with similar motion patterns, based on which local descriptors are constructed for accurate matching. Interest points were accumulated across short image sequences for capturing sufficient appearance variability (Hamdoun et al., 2008). Manifold geometric structures in image sequences of people were utilised to construct more compact spatial descriptors of people (Cong et al., 2009). A histogram of local descriptors based on SIFT (Lowe, 2004) is built from tracks for matching tracked people across view along with an incremental learning (Teixeira and Corte-Real, 2009). The time index of image frames and identity consistency of a sequence were used to constrain spatial feature similarity estimation (Karaman and Bagdanov, 2012). In (Oreifej et al., 2010), a selective region based matching formulation is derived for identity recognition in aerial images, where multiple images of a target are manually labelled. There are also attempts on training a person appearance model from image sets (Nakajima et al., 2003; Bak et al., 2012) or by selecting best pairs (Li and Wang, 2013b). Multiple images of a person sequence were often used either to enhance local image region/patch spatial feature description (Gheissari et al., 2006; Farenzena et al., 2010; Cheng et al., 2011; Xu et al., 2013), or to extract additional appearance information such as appearance change statistics (Bedagkar-Gala and Shah, 2012).

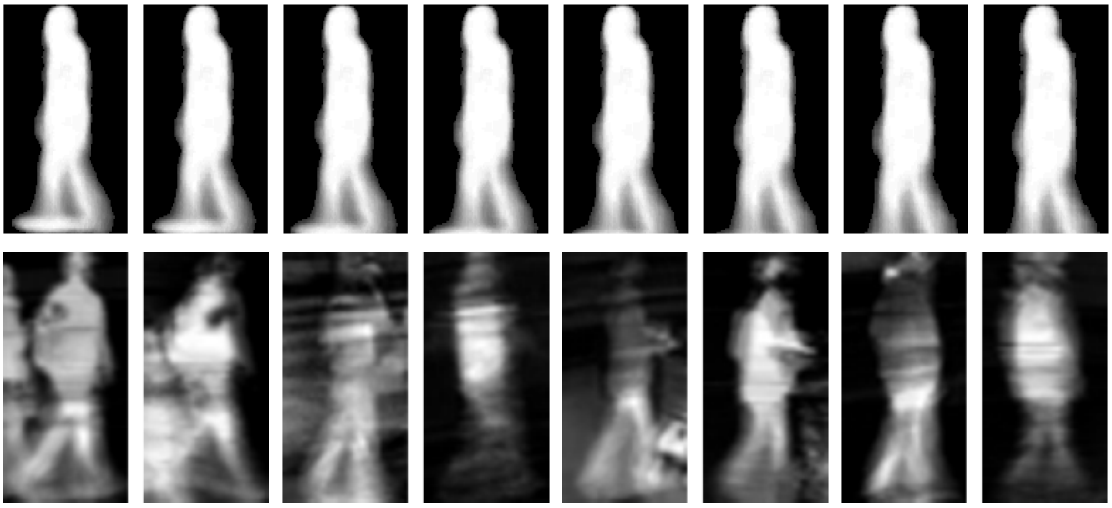


Figure 2.7: Examples of Gait Energy Image (GEI) template computed from image sequences with varying degrees of noise. Top row: GEI templates obtained from some gait recognition data (Han and Bhanu, 2006). Bottom row: GEI features extracted from some person ReID sequences, e.g. the iLIDS-VID dataset (Wang et al., 2014b). It is clearly shown that the ReID image sequences captured in public spaces are much more noisy and challenging than the walking sequence data investigated in gait recognition.

Additionally, other useful information can be potentially extracted and exploited from image sequences of people for helping person ReID, e.g. space-time dynamic features beyond static appearance.

Gait Recognition Space-time information has been explored extensively for gait recognition. Its aim is to develop techniques for person recognition using image sequences by discriminating subtle distinctiveness in the style of walking (Nixon et al., 2010; Sarkar et al., 2005; Han and Bhanu, 2006; Martín-Félez and Xiang, 2012). Gait is a behavioural biometric that measures the way people walk. An advantage of gait recognition is no assumption being made on either subject cooperation (framing) or person distinctive actions (posing). These characteristics are similar to person re-identification situations. However, existing gait recognition models are subject to stringent requirements on person foreground segmentation and accurate alignment over time throughout a gait image sequence or a walking cycle. It is also assumed that complete gait/walking cycles were captured in the target image sequences (Han and Bhanu, 2006; Martín-Félez and Xiang, 2012). Most gait recognition methods do not cope well with cluttered background and/or random occlusions with unknown covariate conditions (Bashir et al., 2010). Person re-identification in public spaces is thus inherently challenging for gait recognition techniques (Figures 1.6 and 2.7). That is, it is challenging to extract a suitable gait representation from

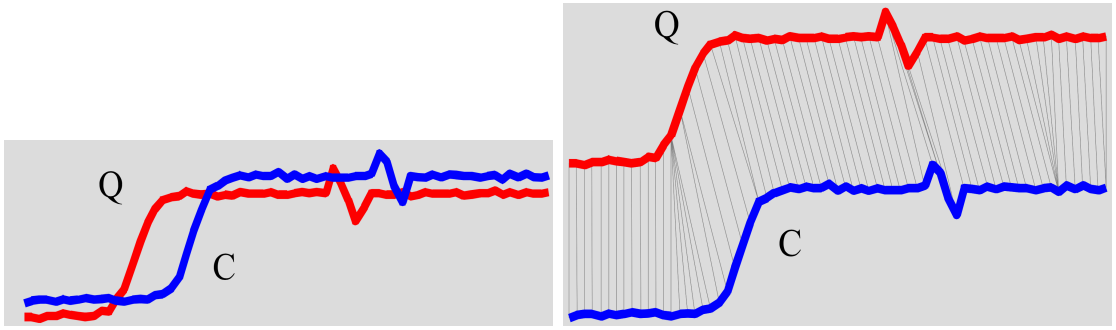


Figure 2.8: An example showing the time warping result (right) between two input time sequences (left) using the DTW algorithm. This figure is reproduced from (Ratanamahatana and Keogh, 2004).

such re-identification data, as shown in our experiments (Section 6.3). In contrast, the approach presented in Chapter 6 relaxes significantly these assumptions by simultaneously selecting discriminative video fragments from noisy image sequences, and matching them cross-view without temporal alignment.

Temporal Sequence Matching An alternative approach to exploiting image sequences for person re-identification is holistic sequence matching. For instance, Dynamic Time Warping (DTW) is a popular sequence matching method widely used for speech recognition (Rabiner and Juang, 1993), action recognition (Lin et al., 2009), and more recently also for person re-identification (Simonnet et al., 2012). The DTW algorithm assumes the alignment at starting and ending data points in the matched sequences, and also the same amount of periodicities, as shown in Figure 2.8. However, given two sequences with unsynchronised starting and/or ending frames, it is difficult to align sequence pairs' starting and ending frames for accurate matching, especially when the image sequences are subject to significant noises caused by unknown camera viewpoint changes, background clutters and drastic lighting changes. The approach presented in Chapter 6 is designed to address this problem so as to avoid any implicit assumptions on sequence alignment and camera view similarity among image frames both within and between sequences.

2.5 Summary

The preceding sections have discussed important studies in the literature with respect to single-camera and multi-camera visual data structure discovery techniques in the generic machine learning and pattern recognition context. Specifically, the main topics include unsupervised and semi-supervised cluster structure analysis, multi-source data clustering, and person identity structure

discovery. Despite the promising results achieved by existing methods, there still exist many limitations and open problems. In the following chapters, novel approaches are presented to overcome the challenges as outlined below:

1. (Chapter 3) **Unsupervised visual data structure discovery with discriminative features**: Unsupervised visual data structure discovery aims to identify the inherent cluster relationships among data samples. This largely helps the understanding and analysis of visual data by proving concise but hidden structural information. Most existing cluster analysis methods typically rely on the whole feature space. Therefore, they are likely to produce sub-optimal results, particularly with high-dimensional and noisy visual features. To address the limitations, a discriminative feature driven clustering framework is formulated.
2. (Chapter 4) **Semi-supervised visual data structure discovery with sparse and imperfect pairwise constraints**: Pairwise constraints provide methods with guidance information to find desired clusters through giving meaningful data similarity clues. In constrained clustering, often sparse constraints thus limited supervision are available. Moreover, some unknown pairwise constraints can be inaccurate. Nonetheless, how to deal with noisy constraints from imperfect oracles is largely ignored in the literature. This thesis presents a constrained clustering model characterised by accurately propagating sparse constraints through discriminative features and the capability of effectively handling noisy constraints.
3. (Chapter 5) **Multi-source data structure discovery for video summarisation**: Multi-source data structure discovery or clustering has not been investigated as extensively as the single source counterpart. Existing algorithms typically treat each individual source data separately and thus ignore the inherent correlations between different sources. This can lead to poor clustering, especially given heterogeneous sources differing significantly in representation, dimension, scale and covariate. To overcome the above issues, a multi-source data structure modelling framework is designed. This method is able to discover and exploit the latent correlations between heterogeneous source data for facilitating the underlying video data structure discovery. The effectiveness of this model is further validated in video summarisation, an important surveillance application that depends greatly on precise video structural information.
4. (Chapter 6) **Person identity structure discovery by video ranking**: Person identity structure or person re-identification across distributed camera views in public spaces are essen-

tial information to video surveillance. The conventional methods mostly exploit people appearance alone to perform cross-view person matching. This is inherently limited due to the large appearance ambiguities and viewing condition disparity between different views. In contrast to existing models, an image sequence based ReID method is developed to extract discriminative space-time features from noisy and unaligned image sequences for achieving more reliable person ReID.

Chapter 3

Unsupervised Visual Data Structure Discovery by Discriminative Features

---

Unsupervised visual data structure discovery or cluster analysis is a fundamental learning strategy and also an essential means of video analysis. The objective is to obtain the underlying data group/cluster membership based on *visual appearance alone*, which however is non-trivial. This is largely because visual signals can be inevitably inaccurate and noisy owing to uncontrollable sources of variation, changes in illumination, random occlusions and background clutters (Gong et al., 2011). More precisely, noisy visual observation with large intra-cluster variations and small inter-cluster differences raises the challenge of accurately measuring the meaningful similarity between data samples, particularly in high-dimensional feature spaces. It is important to overcome this difficulty for visual data cluster analysis, i.e. once meaningful data pairwise similarity is obtained, one can simply construct affinity graphs and then apply any existing graph based clustering algorithms, e.g. spectral clustering (Von Luxburg, 2007; Zelnik-manor and Perona, 2004), to find accurate data clusters. In other words, the performance of graph based clustering methods generally rely significantly on the goodness of the input data affinity graph.

The goal of this chapter is to infer robust pairwise similarity between samples for improving data clustering. Trusting all available features blindly may be susceptible to unreliable and/or noisy features. In light of this, based on unsupervised clustering random forests (Criminisi and Shotton, 2012; Pei et al., 2013; Zhu et al., 2013), a generalised data similarity inference framework is formulated to exploit discriminative features for obtaining more accurate data similarity.

In contrast to confining the notion of similarity to the L_2-norm metric on complex and potentially non-Euclidean behaved data, the proposed model adopts the information-theoretic definition of data similarity as presented in (Lin, 1998).

This chapter is structured as follows. Section 3.1 presents the details of the proposed unsupervised data similarity inference framework based on random forests. This is followed by a description of datasets and experimental settings in Section 3.2. In Section 3.3, the effectiveness of this proposed method was validated by extensive experiments and comparison with state-of-the-art similarity computing models in clustering challenging datasets, including images and surveillance videos. Finally, a summary is presented in Section 3.4.

3.1 Robust Affinity Graph Inference by Discriminative Features

The proposed affinity graph construction approach is built upon conventional clustering random forests, which are an unsupervised form of random forests. The proposed model has a few important merits as below.

1. Our model is purely unsupervised without requiring any ground truth annotations, since it is based on clustering forests rather the more popular supervised classification or regression random forests (Breiman, 2001; Criminisi and Shotton, 2012).
2. By virtue of the random subspace feature selection during training forests, the pairwise affinity matrix generated by our model is less susceptible to corruption of noisy and irrelevant features.
3. Each decision tree in the forest hierarchically encodes an exhaustive set of comparative tests or split functions, which implicitly define different notions of between-sample similarities. Our model is capable of extracting and combining these subtle similarities at distributed discriminative subspaces for learning robust pairwise affinity matrices.

Next, we discuss how to derive robust pairwise similarities from a trained random forest. Recall that the forest training procedure allows us to partition data with very complex distributions at the discovered discriminative feature subspaces. More details on how to train individual decision trees of a conventional forest, and the discriminative feature selection mechanism can be found in Section 2.2.

Specifically, each split function (Equation (2.1)) encodes a different notion of between-sample similarity, defined by its split variable and threshold. To quantify data similarities for gen-

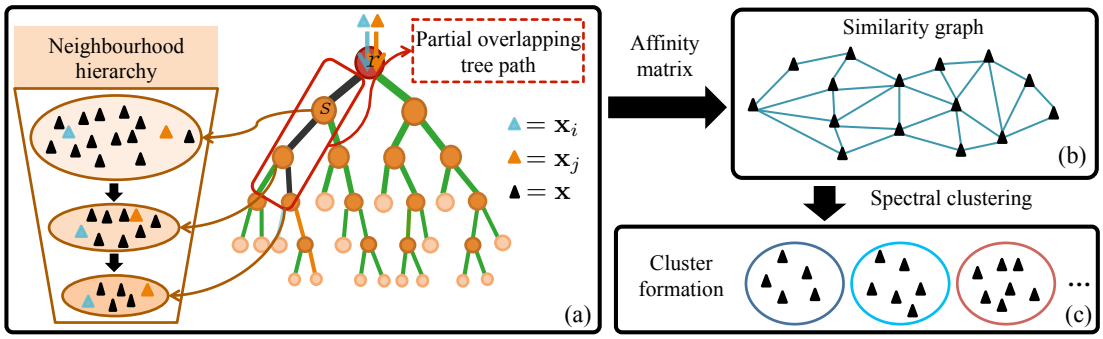


Figure 3.1: Pipeline of visual data structure discovery by clustering, with focus on the hierarchical neighbourhoods along a tree path in a clustering tree, which are formed by selecting and employing discriminative features. The proposed model exploits the hierarchical tree structures and neighbourhoods for robust data pairwise similarity inference.

erating a robust pairwise affinity matrix, we propose a *structure-aware affinity inference model (ClustRF-Strct)* based on clustering random forest. The model takes into account the whole tree hierarchical structures, i.e. a tree path from the root until leaf nodes traversed by data samples $\mathbf{x}$ (Figure 3.1-(a)). Specifically, given the t-th clustering tree, we channel a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ from the root node r until reaching their respective leaf nodes $l(\mathbf{x}_i)$ and $l(\mathbf{x}_j)$. Subsequently, two tree paths composed by the root node r, internal and leaf nodes can be generated:

$$P^i = \{r, s^i_1, \ldots, s^i_k, \ldots, l(\mathbf{x}_i)\}, \tag{3.1}$$
$$P^j = \{r, s^j_1, \ldots, s^j_k, \ldots, l(\mathbf{x}_j)\}, \tag{3.2}$$

with s^i_k and s^j_k denoting the k-th internal nodes travelled by $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.

Intuitively, a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ is considered dissimilar if they are split at the very beginning, e.g. from the root node r. On the other hand, if the samples travel together passing the same set of internal nodes till the identical leaf node, i.e. $P^i = P^j$, their similarity is high. Beyond the two extreme cases above, there exist intermediate similarities: let λ be the length of which P^i and P^j overlaps (Figure 3.1-(a)), i.e.

$$\begin{cases} s^i_k = s^j_k & \text{if } k = \{1, \ldots, \lambda\}, \\ s^i_k \neq s^j_k & \text{if } k = \{\lambda + 1, \ldots\}, \\ l^i \neq l^j. \end{cases} \tag{3.3}$$

Clearly, a larger value in λ signifies more split tests both samples $(\mathbf{x}_i, \mathbf{x}_j)$ have gone through together, implying higher similarity shared between them. A lower value in λ suggests subtle and weak similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. To capture different strengths of data similarities, we derive a principled and generalised tree structure aware data pairwise similarity inference method,

ClustRF-Strct, as

$$a^t_{i,j} = \frac{\sum_{k=1}^{\lambda} w_k}{\sum_{k=1}^{\zeta} w_k}, \tag{3.4}$$

where $\zeta = \max(|P^i|, |P^j|) - 1$, and w_k is the weight assigned to the corresponding tree node (i.e. either s_k or l) on the longer tree path. That is, the longer tree path is utilised as the normalisation factor. Note that the root node r is not considered in computing the similarity since all samples share the same root node. The pairwise similarity $a^t_{i,j}$ defines the individual elements of a tree-level affinity matrix $\mathbf{A}^t \in \mathbb{R}^{n\times n}$, with n the data sample number. To combine consensus from multiple decision trees in the forest, we generate the final smooth affinity matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ as

$$\mathbf{A} = \frac{1}{\tau_{\text{clust}}} \sum_{t=1}^{\tau_{\text{clust}}} \mathbf{A}^t. \tag{3.5}$$

Note that the Equation (3.5) is adopted as the ensemble model of random forest due to its advantage of suppressing the noisy tree predictions, though other alternatives such as the product of tree-level predictions are possible (Criminisi and Shotton, 2012).

ClustRF-Strct is regarded as a generic affinity inference model since distinct strategies of defining node weights w_i can produce different affinity graph construction methods/instantiations, as we will describe below.

3.1.1 Variant I - The Binary Affinity Model

We show that the methods proposed in (Criminisi and Shotton, 2012; Pei et al., 2013; Zhu et al., 2013) are special cases of the proposed ClustRF-Strct. All these methods share the same mechanism in estimating a pairwise similarity matrix using a clustering random forest. We name these methods collectively as *the binary affinity inference model (ClustRF-Bi)*, since they derive pairwise affinity based only on whether or not (binary) two samples fall into the same leaf node of a tree.

Prior to discussing their relationship to our approach, we review the underlying mechanism of ClustRF-Bi in measuring pairwise similarity between data samples given a learned clustering forest. Recall that each individual tree of a forest partitions the training samples at its leaves $\{l(\mathbf{x})\}$ where $l(\mathbf{x})$ represents a leaf node $\mathbf{x}$ falls into in a given tree. For each tree, the ClustRF-Bi model first computes a tree-level $n \times n$ affinity matrix $\mathbf{A}^t$ with elements defined as

$$a^t_{i,j} = \exp^{-\text{dist}^t(\mathbf{x}_i, \mathbf{x}_j)}, \text{ with} \tag{3.6}$$

$$\text{dist}^t(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0, & \text{if } l_t(\mathbf{x}_i) = l_t(\mathbf{x}_j), \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.7}$$

With Equation (3.6), the ClustRF-Bi assigns the maximal similarity $a^t_{i,j} = 1$ to a sample pair ($\mathbf{x}_i$, $\mathbf{x}_j$) if $P^i = P^j$ (i.e. completely overlapping), and the minimum similarity $a^t_{i,j} = 0$ to them otherwise, regardless of any partial overlap in their tree paths. This formulation is equivalent to setting $w_k = 0$ for every internal node, $w_k = 1$ for all leaf nodes in Equation (3.4). Hence, this mechanism is a special case of our ClustRF-Strct. A potential problem with ClustRF-Bi is that it may lose the weak and subtle proximity of sample pairs proportional to the degree of path overlap. We will show in our experiments in Section 3.3 that considering only completely overlapping path pairs, i.e. $P^i \backslash P^j = \emptyset$, as in ClustRF-Bi, is not sufficient for producing satisfactory data clusters.

3.1.2 Variant II - The Uniform Structure Model

To address the limitation of ClustRF-Bi in losing weak similarity between data samples, we propose to consider the non-completely-overlapping path pairs as well while measuring tree-level data similarities using the proposed ClustRF-Strct model. In particular, we treat all tree nodes as uniformly important by setting $w_k = 1$ in Equation (3.4). Therefore, Equation (3.4) can be rewritten as

$$a^t_{i,j} = \frac{\lambda}{\max(|P^i|, |P^j|) - 1}. \tag{3.8}$$

We call this model as *ClustRF-Strct-Unfm*. With Equation (3.8), all partially overlapped path pairs also contribute to the similarity estimation between samples. As shown in the experiments (Section 3.3), this formulation captures weak data similarities encoded in the tree structures, and thus is capable of better revealing the underlying data structure than the conventional ClustRF-Bi model.

3.1.3 Variant III - The Adaptive Structure Model

The ClustRF-Strct-Unfm is capable of capturing subtle and weak data proximity through exploiting the path sharing mechanism of sample pairs in the hierarchical structure of the forest. Nevertheless, the uniform node weighting implies an implicit assumption that all tree nodes (e.g. s_k or l) are equally important in defining similarity. In reality this may not be true, particularly with data of complex distributions, since different nodes reside at distinct layers of the tree hierarchy with dissimilar properties, e.g. the size and structure of the arrival training samples.

To characterise such node (or data subset) properties, we propose an *adaptive structure-aware affinity inference (ClustRF-Strct-Adpt)*.

The ClustRF-Strct-Adpt model exploits the *hierarchical neighbourhood* formed in each clustering tree (see Figure 3.1-(a)). Our notion of hierarchical neighbourhood generalises the idea presented in (Lin and Jeon, 2002). Specifically, (Lin and Jeon, 2002) only regards samples sharing the same tree terminal node as neighbours. We extend the neighbourhood notion to the whole tree hierarchy. Imagine a situation where a target sample $\mathbf{x}_t$ traverses in a tree hierarchy from the root node until some arbitrary internal node s_k. Some other samples $S_k \setminus \mathbf{x}_t$ have also gone through the same tree path and fall onto the same internal node s_k with $\mathbf{x}_t$. These samples form a neighbourhood with $\mathbf{x}_t$ on node s_k in the tree hierarchy.

Samples that form a hierarchical neighbourhood have passed through the same set of split functions (Equation (2.1)) associated with each tree node. Intuitively, the deeper the hierarchical neighbourhood is formed, the higher the similarity shared among the samples in the same neighbourhood, since those samples have survived and are still connected after identical discriminative split tests (Equation (2.1)). Motivated by this observation, we assign each tree node s_k with a scale-adaptive weight (Equation (3.4)) as

$$w_k = \frac{1}{|S_k|}. \tag{3.9}$$

Consequently, we assign larger weights to deeper tree nodes, since $|S_k| > |S_{k+1}|$. As such, ClustRF-Strct-Adpt estimates similarity between a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ via

$$a^t_{i,j} = \frac{\sum_{k=1}^{\lambda} \left(\frac{1}{|S_k|} \right)}{\sum_{k=1}^{\zeta} \left(\frac{1}{|S_k|} \right) + \frac{1}{|B|}}, \tag{3.10}$$

where B denotes the set of data samples reaching into the leaf node l^b of the longer path. Similar to Equation (3.8), a maximum similarity is assigned to sample pairs that share the same leaf node. Nevertheless, the tree node similarity weight is no longer distributed linearly along the forest hierarchy as in Equation (3.8), but in a non-linear way adaptive to the size of hierarchical neighbourhood.

3.2 Datasets and Experimental Settings

Datasets A variety of visual datasets were utilised for evaluating the proposed model: (1) Image Segmentation (Asuncion and Newman, 2007): a scene image dataset from the UCI repository, including 7 types of different outdoor scenes: Brickface, Sky, Foliage, Cement, Window,

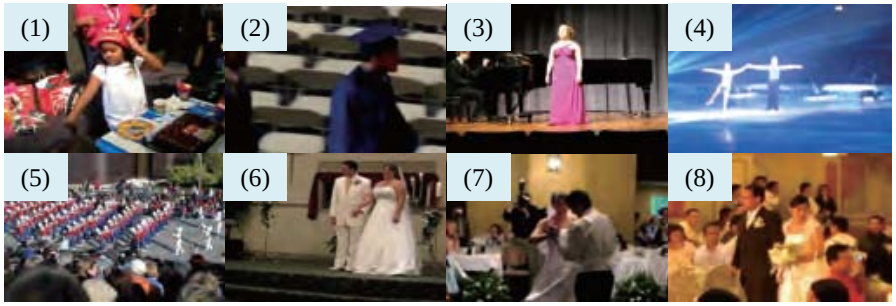
Table 3.1: Dataset statistics, with examples in Figure 3.2.

Dataset	# Clusters	# Features	# Samples
Image Segmentation (Asuncion and Newman, 2007)	7	19	2310
CMU-PIE (Sim et al., 2003)	10	1024	1000
USAA (Fu et al., 2014)	8	14000	1466
ERCe (Zhu et al., 2013)	6	2672	600

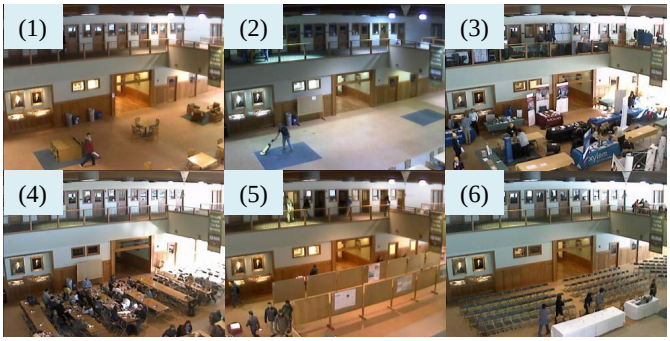
Path, and Grass. The objective is to partition image patches into the above seven types. (2) CMU-PIE (Sim et al., 2003): a face image dataset drawn from CMU-PIE. It comprises 10 different persons selected in random, each with 100 images of near frontal poses and various expressions and lighting conditions (Figure 3.2a). We aim to group together all the face images from the same person on this dataset. (3) USAA (Fu et al., 2014): a YouTube video dataset. This dataset features common social group activities where unconstrained space of objects, events and interactions makes them intrinsically complex and challenging to detect (Figure 3.2b). The goal is to cluster these video clips into 8 groups each with coherent semantics, e.g. the same social activity. (4) ERCe (Zhu et al., 2013): a visual surveillance video dataset. The dataset is challenging because of various types of physical events characterised by large changes in the environmental set-up, participants, and crowdedness, as well as intricate activity patterns. This dataset consists of 600 video clips from 6 campus events, each with 100 samples (Figure 3.2c). Our purpose is to classify the ERCe video clips into the six events.

Data feature representation For Image Segmentation, USAA, and ERCe, we use the same features as provided by (Asuncion and Newman, 2007), (Fu et al., 2014) and (Zhu et al., 2013). Specifically, for Image Segmentation, we use the low-level visual features from image patches, e.g. colour, pixel intensity. These appearance features may be unreliable and noisy, especially given outdoor scenes. As to USAA, the resulting high-dimensional (14000-D) feature vectors are drawn from three heterogeneous modalities, namely static appearance, motion and auditory. The data samples from ERCe are also of high-dimensional (2672-D), involving heterogeneous feature types, e.g. colour histogram (RGB and HSV), optical flow, local texture, holistic image appearance, object detection. With CMU-PIE, we first normalise and crop the face images into 32×32 in spatial resolution, and their raw pixel values are then employed as the representation. Such a representation is affected by large differences in illumination, facial expression, and head

(a) CMU-PIE (Sim et al., 2003): each row corresponds to one person.



(b) USAA (Fu et al., 2014): (1) Birthday Party, (2) Graduation, (3) Music Performance, (4) Non-music Performance, (5) Parade, (6) Wedding Ceremony, (7) Wedding Dance, (8) Wedding Reception.



(c) ERCe (Zhu et al., 2013): (1) Student Orientation, (2) Cleaning, (3) Career Fair, (4) Group Study, (5) Gun Forum, (6) Scholarship Competition.

Figure 3.2: Example images from CMU-PIE (Sim et al., 2003), USAA (Fu et al., 2014), ERCe (Zhu et al., 2013) datasets.

pose. All data features are scaled to the range of $[-1,1]$. To initially remove less-informative features on the high-dimensional datasets, e.g. CMU-PIE, USAA and ERCe, we perform PCA on them and the first 30 dominant components are used as the final representation. The same sets of feature data are used across all methods for fair comparison.

Baselines We compare the proposed affinity graph learning model ClustRF-Strct with:

1. *k Nearest Neighbours (kNN)* (Wang et al., 2008): the most traditional affinity graph construction method using the Euclidean distance on the input feature space. To convert an Euclidean distance matrix $\mathbf{D}$ into an affinity graph $\mathbf{A}$, we compute each element in $\mathbf{A}$ as $a_{i,j} = \exp(-\text{dist}^2_{i,j}/\sigma^2_{i,j})$ with σ_{ij} the adaptive kernel size that is computed as the mean distance of k_{adpt}-nearest neighbourhoods as in (Wang et al., 2008). We will evaluate the sensitivity of k_{adpt} on the clustering performance in Section 3.3.
2. *Dominant Neighbourhoods (DN)* (Pavan and Pelillo, 2007): a tight affinity graph learning approach. To reduce the amount of potentially noisy edges in a given Euclidean affinity graph, the DN model attempts to identify sparse and compact neighbourhoods through selecting only the maximal cliques in the input graph.
3. *Consensus of kNN (cons-kNN)* (Premachandran and Kakarala, 2013): the state-of-the-art affinity graph construction method. For selecting strong local neighbourhoods, the consensus information collected from various neighbourhoods in a provided kNN graph is exploited by this algorithm for producing a more robust affinity graph.
4. *ClustRF-Bi* (Criminisi and Shotton, 2012; Pei et al., 2013; Zhu et al., 2013): the clustering random forest binary affinity model (Section 3.1.1). This method exploits discriminative features identified during the training of clustering forests to construct data affinity graphs. The resulting affinity graphs can thus be less-sensitive to noisy features, compared to the Euclidean-metric-based methods, e.g. kNN, DN and cons-kNN.

Evaluation metrics We use the widely adopted adjusted Rand Index (ARI) (Hubert and Arabie, 1985) as the evaluation metric, with the range of $[-1,1]$. ARI measures the agreement between the clustering results and the ground truth in a pairwise fashion, with higher values indicating better clustering quality. ARI assumes the generalised hypergeometric distribution on models (e.g. partitions or clustering results), with the general index form as:

$$i_{\text{ghd}} = \frac{\hat{m} - e}{m - e} \tag{3.11}$$

where $\hat{m}$, m and e refer to the actual index, maximal index, and expected index, separately. Therefore, the expected value of i_{ghd} is 0. We formally define the ARI below. Let $U = \{U_1, U_2, \ldots, U_r\}$ and $V = \{V_1, V_2, \ldots, V_s\}$ be the ground truth and clustering result partition of n data samples, respectively. We then denote m_{ij} as the number of data samples that are in U_i and V_j, $m_{i\cdot}$ and $m_{\cdot j}$ as the number of sample in the cluster U_i and V_j. The ARI is computed as:

$$\text{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}} \tag{3.12}$$

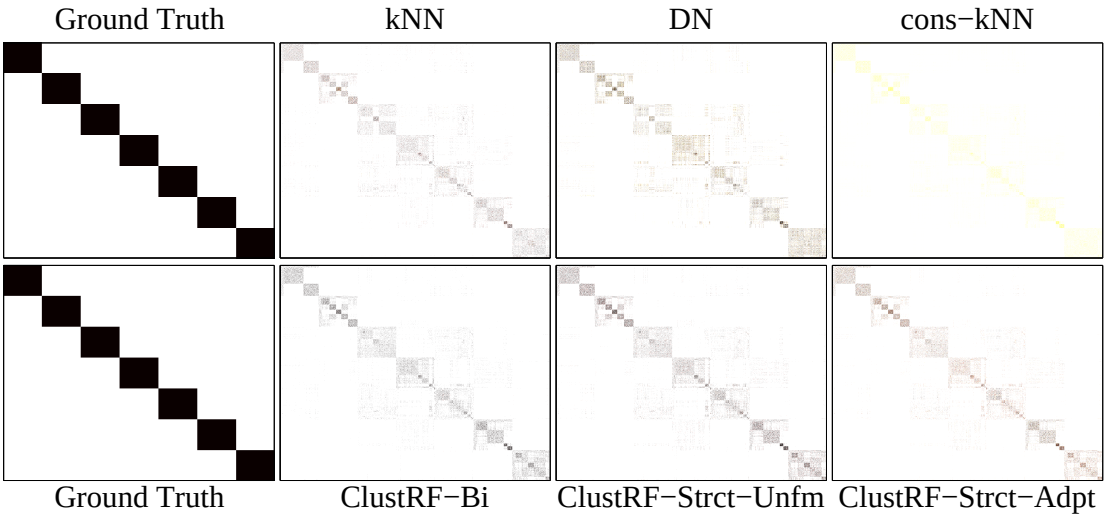
The quantity $\sum_{i,j}\binom{n_{ij}}{2}$ (e.g. the actual index) means the number of sample pairs that are in the same cluster in both U and V, which can be regarded as agreement between the two partitions; $\left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}$ (e.g. the expected index) measures the expected number of sample pairs in the same cluster; and $\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right]$ (e.g. the maximal index) refers to the summed same-cluster pairs in U and V. As a result, a higher ARI value means greater agreement of the clustering result V with the ground truth partition U, i.e. more agreed same-cluster data pairs w.r.t. their expected value. Consider two clustering result partitions V^a and V^b with the same cluster number, if their $\sum_j \binom{n_{\cdot j}}{2}$ is the same, then comparing their ARI values can directly reflect their disparity w.r.t. the number of same-cluster pairs that are consistent with U; Otherwise, approximately reflected. In our evaluation, for all experiments involving clustering forest based models, i.e. ClustRF-Bi, ClustRF-Strct-Unfm, and ClustRF-Strct-Adpt, we report the ARI values averaged over 5 trials.

Implementation details The number of trees τ_{clust} in a clustering forest is set to 1000. We observed stable results given a larger forest size. This observation agrees with (Criminisi and Shotton, 2012). We set d_{try} (see Equation (2.2)) to $\sqrt{d}$ with d the feature dimensionality of the input data and employ a axis-aligned data separation hyperplane (Criminisi and Shotton, 2012) as the split function (see Equation (2.1)). The value of ϕ is obtained through cross-validation on each dataset.

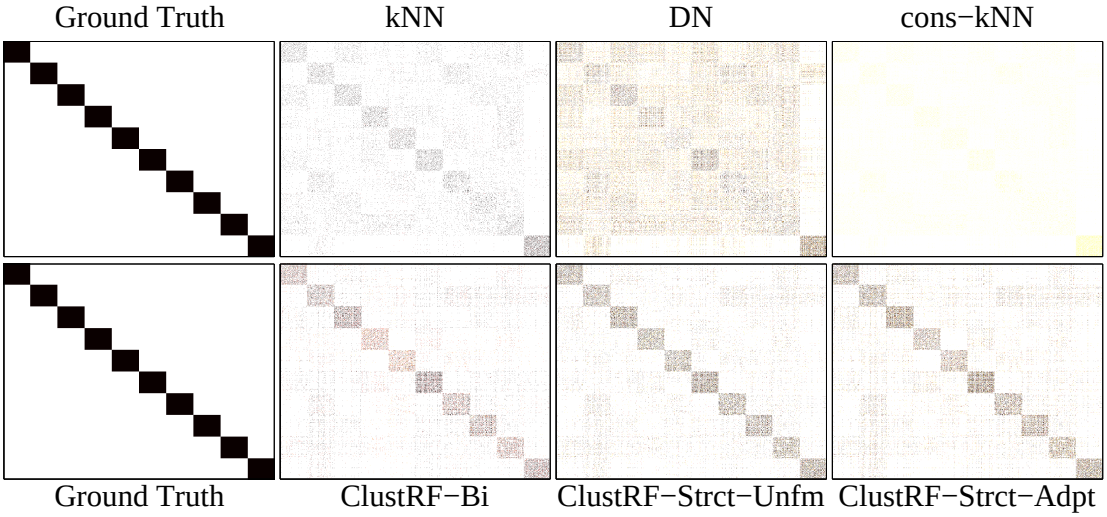
3.3 Experiments and Evaluations

3.3.1 Evaluation on Affinity Graph

We first examine the data affinity graphs, which could qualitatively reflect how effective a neighbourhood graph construction method is. Figure 3.3 depicts some example affinity matrices generated by all comparative models.



(a) Image Segmentation (Asuncion and Newman, 2007)



(b) CMU-PIE (Sim et al., 2003)

Figure 3.3: Qualitative comparison of the affinity graphs generated by different methods.

It can be observed that ClustRF-Strct-Unfm and ClustRF-Strct-Adpt produce affinity matrices with more distinct block structure and less false edges compared with others. This suggests the superiority of the proposed models in learning the underlying structures of data, potentially leading to more compact and separable clusters. A number of noisy pairwise edges are found in the affinity graphs yielded by ClustRF-Strct-Unfm than those by ClustRF-Strct-Adpt. This is a consequence of not considering the goodness of hierarchical neighbourhoods in ClustRF-Strct-Unfm (Section 3.1.2), leading to less accurate induced data similarities in comparison to ClustRF-Strct-Adpt. This observation shows the effectiveness of the proposed adaptive weighting mechanism in suppressing noisy or inaccurate features on learning data sample proximity.

We now examine and discuss the characteristics of affinity matrices constructed by other

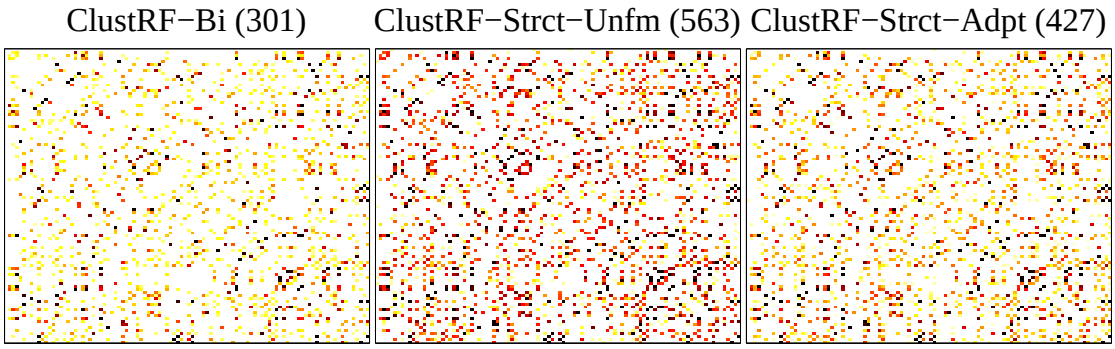


Figure 3.4: Comparison between clustering forest based models: the pairwise affinity between different face images from the same person (CMU-PIE (Sim et al., 2003)). The numbers in the parentheses are the summation of all pairwise similarities induced by the corresponding method. Larger is better.

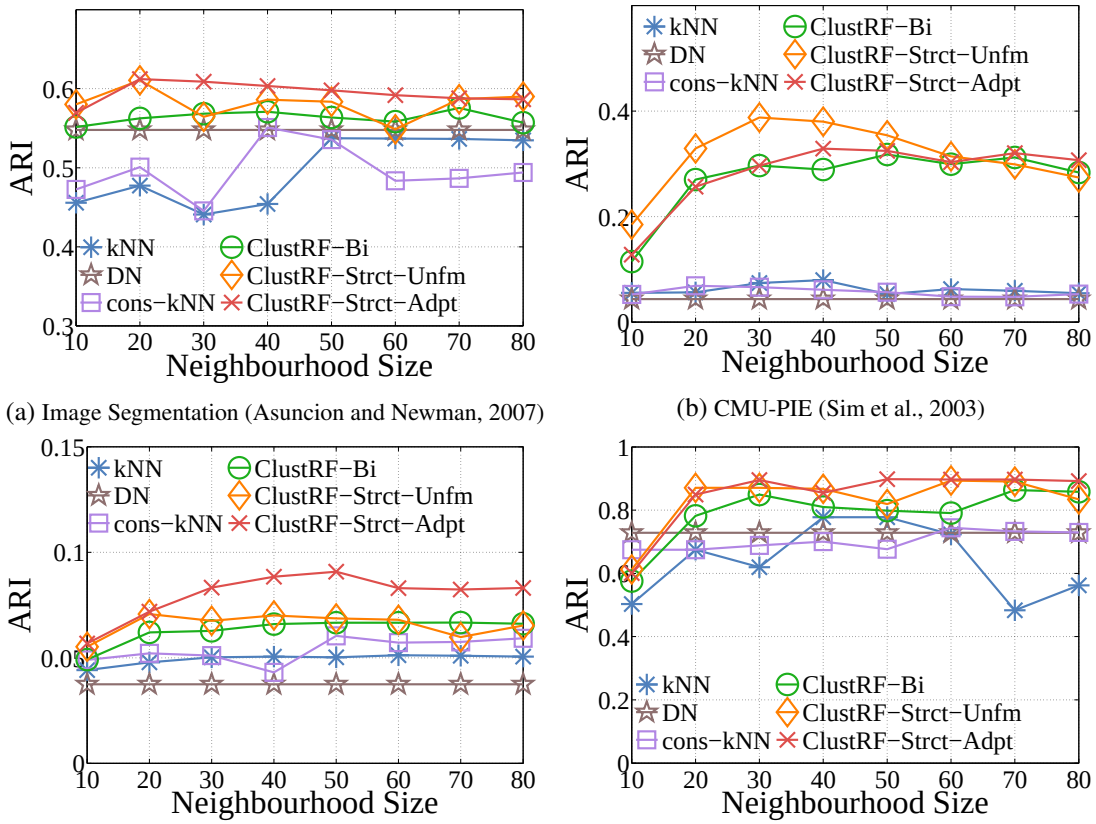
baselines. It is observed from Figure 3.4 that compared to the ClustRF-Strct models, ClustRF-Bi has the tendency to underestimate the similarity of sample pairs that actually originate from the same clusters. This is owing to that ClustRF-Bi only assumes data similarity on the completely overlapped tree path pairs, and thus loses subtle and weak data proximity (Section 3.1.1). Given intrinsically ambiguous datasets with unreliable features, incomplete overlapping path pairs can often occur as samples of the same categories may only share similarity in some feature subspaces. In such cases, ClustRF-Bi shall perform poorly as compared to our ClustRF-Strct models, as we shall show next.

With *k*NN, DN, and cons-*k*NN, affinity graphs with indistinct block structure are observed, with a mix of large quantity of faulty edges. In contrast to ClustRF-Bi that is 'overly reluctant' in assigning data proximity to sample pairs, the Euclidean distance based methods go to the other extreme by blindly believing all available features and therefore tend to introduce false data proximity.

3.3.2 Evaluation on Data Structure Discovery

In this experiment, we quantitatively evaluate data clustering performance of different graph construction methods by applying the spectral clustering algorithm (Zelnik-manor and Perona, 2004) on their affinity graphs as discussed in Section 3.3.1.

It is observed from Figure 3.5 and Table 3.2 ClustRF-Strct-Unfm and ClustRF-Strct-Adpt outperform baseline methods, e.g. by as much as $>$**125**% and $>$**120**% relative improvement against *k*NN, $>$**190**% and $>$**180**% against DN, $>$**130**% and $>$**125**% against the state-of-the-art cons-*k*NN, $>$**5**% and $>$**10**% against the discriminative-feature-based model ClustRF-Bi in



(a) Image Segmentation (Asuncion and Newman, 2007)

(b) CMU-PIE (Sim et al., 2003)

(c) USAA (Fu et al., 2014)

(d) ERCe (Zhu et al., 2013)

Figure 3.5: ARI score against neighbourhood size: comparison between different methods on the spectral clustering performance given different scales of neighbourhood k. The neighbourhood size k_{adpt} used on computing the adaptive Gaussian kernel size is fixed to 20.

terms of the area under the curve of ARI score against neighbourhood size averaged over all the datasets. This is in line with the observations in Figure 3.3. Importantly, we find that ClustRF-Strct-Unfm and ClustRF-Strct-Adpt significantly outperform the Euclidean distance based methods on CMU-PIE. This can be due to the capability of our model of capturing and aggregating subtle data proximity distributed over discriminative feature subspaces, thus suitable to handle ambiguous and unreliable features caused by variation in illumination, face expression or pose on the CMU-PIE data. A large improvement margin is also observed on the USAA dataset with data collected from heterogeneous sources. All these evidences suggest the superior capability of our model in dealing with high-dimensional data and heterogeneous sources for generating robust affinity graphs. Furthermore, ClustRF-Strct-Adpt is superior to ClustRF-Strct-Unfm in most cases except on the CMU-PIE dataset. The plausible reason is that: the feature importance of face images is relatively more uniform due to little background observation, compared to USAA and ERCe where more distractive background clutter is involved, and the adaptive weight thus somewhat violates this uniform importance pattern.

As shown in Figure 3.5, ClustRF-Strct-Unfm is more likely to suffer when the size of neighbourhood k increases, whilst ClustRF-Strct-Adpt behaves more stably. The tendency is likely to be caused by the relatively noisier affinity matrix induced by ClustRF-Strct-Unfm, as we observed in Section 3.3.1. The results further justify the importance of considering neighbourhood-scale-adaptive weighting on tree nodes (Section 3.1) for suppressing data noise.

The Euclidean-distance-based models produce the poorest results over all the datasets. Inaccurate and noisy features are potential causes. For example, the face images from the CMU-PIE dataset are intrinsically ambiguous owing to large variations in illumination and expressions (Figure 3.2-(a)). The extracted features are therefore unreliable. Similar situations are observed on other datasets. The cons-kNN model attempts to circumvent this problem via searching for consensus from multiple kNNs. Nevertheless this is proved challenging, particularly when a large quantity of potential noisy edges exist in the given kNN due to the unreliable input data, leading to possibly inconsistent neighbour votes from multiple kNNs. DN is likely to suffer from the same problem as the maximal cliques in the given affinity graph is no longer trustworthy. This interpretation is further supported by the fact that for all kNN, cons-kNN and DN, the clustering performance changes dramatically with the varying settings of neighbourhood size k, e.g. on Image Segmentation and ERCe. That is, a large amount of inaccurate edges in the affinity graphs

Table 3.2: Sensitivity of k_{adpt}: the clustering results of different methods given varying values of k_{adpt} in terms of AUC, with k_{adpt} the parameter used for computing the adaptive Gaussian kernel size during the process of converting a Euclidean distance matrix into an affinity graph (see Section 3.2).

Dataset	Image Segmentation					CMU-PIE					USAA					ERCe				
k_{adpt}	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
kNN	34.8	36.2	37.6	37.8	37.9	4.4	4.4	4.9	4.8	4.7	3.5	3.1	3.3	3.6	3.6	45.9	48.1	52.1	52.7	51.8
DN	38.3	29.1	34.7	37.2	37.2	3.0	2.3	2.4	3.0	3.5	2.6	2.3	2.5	2.0	1.7	51.0	52.1	49.9	18.3	25.6
cons-kNN	34.9	36.8	35.8	36.8	35.9	4.0	4.4	4.3	4.3	4.2	3.8	3.8	3.8	3.8	3.9	49.2	52.1	52.0	52.0	55.7
ClustRF-Bi	39.5					19.8					4.5					56.1				
ClustRF-Strct-Unfm	40.7					**22.9**					4.7					59.3				
ClustRF-Strct-Adpt	**41.8**					20.5					**5.7**					**60.4**				

lead to the requirement of a more careful neighbourhood size selection, so as to trade-off between the true and false data similarities.

By exploiting discriminative features, the ClustRF-Bi model suffers less from noisy data, and produces better results than the Euclidean-distance-based methods. However, it is inferior to the proposed ClustRF-Strct variants, since it is not capable of capturing subtle data pairwise similarity encoded in partially overlapped path pairs.

Sensitivity of k_{adpt} Here we evaluate the sensitivity of k_{adpt} on kNN, DN and cons-kNN. The parameter k_{adpt} is employed to estimate the adaptive Gaussian kernel size for converting a Euclidean distance matrix into a similarity graph (Wang et al., 2008) (Section 3.2). Note that ClustRF-Bi, ClustRF-Strct-Unfm and ClustRF-Strct-Adpt are free from k_{adpt} since they directly derive affinity graphs from the learned forests, rather than from distance matrices which require a Gaussian kernel to enforce locality. It is evident from Table 3.2 that for all the Euclidean-distance-based affinity graph learning models, a careful selection of adaptive Gaussian kernel size can produce better clustering results. However, their best results are still worse than those by clustering forest based models, due to the limitation in handling intrinsically noisy and irrelevant feature data. Importantly, the proposed ClustRF-Strct model gains superior performance to other baselines in all cases.

3.4 Summary

This chapter has presented a generic unsupervised approach to constructing more robust and meaningful data affinity graphs for improving unsupervised visual data structure discovery. This

method is designed particularly to deal with the challenges in clustering high-dimensional and heterogeneous visual data with potentially large diversity within clusters and much similarity between clusters caused by non-relative and noisy features. Specifically, instead of blindly trusting all available variables, we adopt an information-theoretic definition for measuring data similarity and quantify affinity degrees through capturing and combining subtle/weak data proximity distributed in discriminative feature subspaces identified during the training process of clustering random forests. Moreover, affinity graphs constructed by the proposed model naturally possess the local neighbourhood, with no need of Gaussian kernel. Extensive experiments on clustering challenging visual datasets have demonstrated the superiority of the proposed affinity inference model over the state-of-the-art models.

Inherently, unsupervised visual data clustering is not well defined and its performance can largely depend on data feature representation in addition to clustering methods. To obtain satisfactory cluster results, prior information e.g. from human experts may be needed and useful to provide some guidance for the clustering behaviour. The next chapter deals with the problem of visual data cluster structure discovery by taking into account additional sample-level supervision during the clustering procedure.

Chapter 4

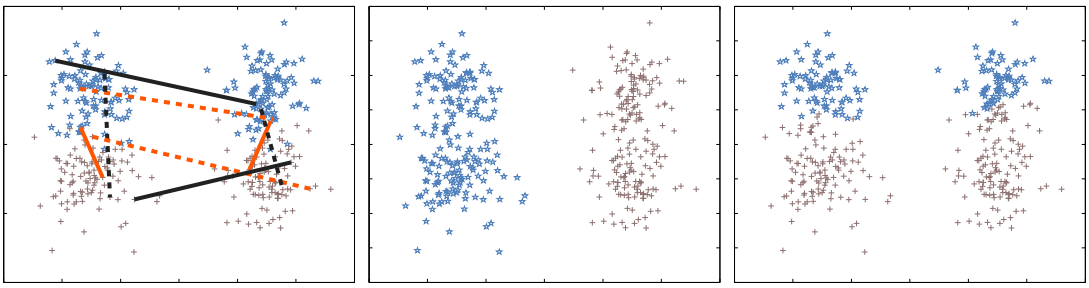
Semi-Supervised Visual Data Structure Discovery with Sparse and Imperfect Pairwise Relationships

---

Unsupervised structure discovery (or clustering) over visual data is somewhat 'blind' and less guided, with heavy dependence on visual feature representations in terms of information source. It is thus a ill-posed problem, i.e. the similarity metric is not explicitly defined, particularly so when only unreliable feature data are available from visual data, mostly typical in video surveillance (Gong and Xiang, 2011). It makes sense to seek and exploit other knowledge when available for obtaining more accurate and desired cluster structure formation.

In clustering context, prior knowledge is typically expressed in form of pairwise constraints or relationships, namely *must-link* - a pair of samples must be in the same cluster, and *cannot-link* - a pair of samples belong to different clusters. As shown in (Wagstaff et al., 2001; Xing et al., 2002; Basu et al., 2004a), clustering data by using such prior belief as constraints in addition to data features (a.k.a. constrained clustering) to influence the cluster formation process helps.

In this chapter, we consider the problem of pairwise similarity based semi-supervised video data structure discovery (or constrained clustering), given pairwise constraints derived from human/oracles (see Figure 4.1). Specifically, constraints together with data feature representations are exploited for helping compute the similarity between data samples so that the induced data neighbourhood relations are more meaningful and expressing the desired high-level structures. Similar to unsupervised video clustering in Chapter 3, the estimated similarity matrix can be utilised to benefit graph based clustering algorithms for inducing more precise data clusters.



(a) Ground Truth (b) Spectral Clustering (c) Proposed Approach

Figure 4.1: (a) Ground truth cluster formation, with invalid pairwise constraints highlighted in light red colour; must- and cannot-links are represented by solid and dashed lines respectively; (b) the result obtained using unsupervised clustering; (c) the clustering result obtained using the proposed method.

The objective of this chapter is to effectively exploit the available pairwise data relationships for helping the revelation of inherent visual data group structure. This is no-trivial because: (1) Often, pairwise constraints are available only in a small quantity and thus the information provided can be very limited; (2) Moreover, these constraints are not guaranteed to be absolutely accurate in reality, and thus may mislead the clustering process.

This chapter is organised as below. Problem formulation and the proposed semi-supervised clustering model are detailed in Section 4.1. Followed by experimental settings in Section 4.2, experiments with extensive evaluations by comparing to a wide range of state-of-the-art semi-supervised clustering methods are presented in Section 4.3. Section 4.4 summarises this chapter.

4.1 Semi-Supervised Visual Data Structure Discovery with Imperfect Oracles

4.1.1 Problem Definition

Given a set of samples denoted as $X = \{\mathbf{x}_i\}$, $i = 1,\ldots,n$, with n denoting the total number of samples, and $\mathbf{x}_i = (x_{i,1},\ldots,x_{i,d})$, d the feature dimensionality of the feature space $F \subset \mathbb{R}^d$, the goal of unsupervised clustering is to assign each sample $\mathbf{x}_i$ with a cluster label c_i. In constrained clustering, additional pairwise constraints are available to influence the cluster formation. There are two typical types of pairwise constraints:

$$\begin{aligned} \text{Must-link} \quad &: \quad ML = \{(\mathbf{x}_i,\mathbf{x}_j) \mid c_i = c_j\}, \\ \text{Cannot-link} \quad &: \quad CL \ = \{(\mathbf{x}_i,\mathbf{x}_j) \mid c_i \neq c_j\}. \end{aligned} \tag{4.1}$$

We denote the full constraint set as $FL = ML \cup CL$. The pairwise constraints may arise from pairwise similarity as perceived by a human annotator (oracle), temporal continuity, or prior

knowledge on the sample class label. Acquiring pairwise constraints from a human annotator is expensive. In addition, owing to data ambiguity and human unintentional mistakes, the pairwise constraints are likely to be incorrect and inconsistent with the underlying data distribution.

We propose a model that can flexibly generate constraint-aware affinity matrices, which can be directly employed as input by existing pairwise similarity based clustering algorithms e.g. spectral clustering (Von Luxburg, 2007) or affinity propagation (Frey and Dueck, 2007) for semi-supervised (constrained) clustering (Figure 4.2). The details of the proposed model are described next.

4.1.2 Constraint Propagation Random Forest

To address the issues of sparse and noisy constraints, we formulate a COnstraint Propagation Random Forest (COP-RF), a variant of clustering forest (see Figure 4.2). We consider using a random forest (Section 2.2), particularly a clustering forest (Breiman, 2001; Zhu et al., 2014; Liu et al., 2000; Blockeel et al., 1998) as the basis to derive our model for two main reasons:

1. It has been shown that random forest has a close connection with adaptive k-nearest neighbour methods, as a forest model adapts neighbourhood shape according to the local importance of different input variables (Lin and Jeon, 2002). This motivates us to exploit the adaptive neighbourhood shape[1] for effective constraint propagation.
2. The forest model also offers an implicit feature selection mechanism that allows more accurate constraint propagation in the provided feature space by exploiting identified discriminative features during model training.

The proposed COP-RF differs significantly from the conventional random forests in that the COP-RF is formulated with a new split function, which considers not only the bottom-up data feature information gain maximisation, but also the joint satisfaction of top-down pairwise constraints. In what follows, we first detail the training of COP-RF followed by how COP-RF performs constraint propagation through discriminative feature subspaces.

Training of COP-RF The training of a COP-RF involves independently growing an ensemble of τ_{clust} constraint-aware COP-trees. To train a COP-tree, we iteratively optimise the split function (Equation (2.1)) by finding the optimal $\hat{\boldsymbol{\vartheta}}$ including both the best feature dimension and cut-point to partition the node training samples S, similar to an ordinary decision tree (Section 2.2).

[1]The neighbours of a data $\mathbf{x}$ in forest interpretation are the points that fall into the same child node.

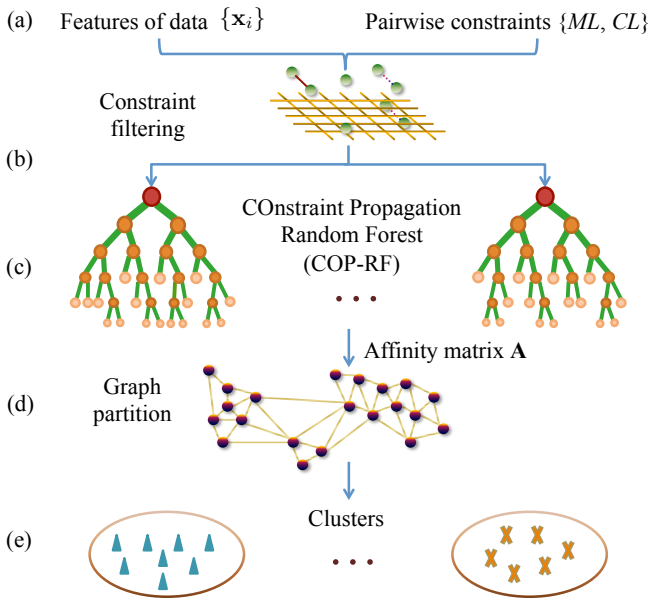


Figure 4.2: Overview of the proposed constrained clustering approach. (a) The inputs into a constrained clustering model: features of data and pairwise constraints; (b) The proposed COP-RF model; (c) Performing data clustering on the derived similarity graph; (d) The obtained cluster formation.

The difference is that the term 'best' or 'optimal' is no longer defined only as to maximising the bottom-up feature information gain, but also simultaneously satisfying the imposed top-down pairwise constraints. More precisely, at the t-th COP-tree, its training set X^t only encompasses a subset of the full constraint set FL, i.e.

$$FL^t = \left\{ML^t \cup CL^t\right\} \subset FL. \tag{4.2}$$

where ML and CL are defined in Equation (4.1). Instead of directly using the information gain in Equation (2.3), we optimise each internal node s in a COP-tree via enforcing additional conditions on the candidate data splits:

$$\begin{aligned} \forall(\mathbf{x}_i,\mathbf{x}_j) \in ML^t \Rightarrow \mathbf{x}_i,\mathbf{x}_j \in L \text{ (or } \mathbf{x}_i,\mathbf{x}_j \in R), \\ \exists(\mathbf{x}_i,\mathbf{x}_j) \in CL^t \Rightarrow \mathbf{x}_i \in L \,\&\, \mathbf{x}_j \in R \text{ (or opposite)}, \\ \text{where } \mathbf{x}_i,\mathbf{x}_j \in S, \text{ and } FL^t = ML^t \cup CL^t. \end{aligned} \tag{4.3}$$

L and R are data subsets at left and right child (see Equation (2.3)). Owing to the conditions in Equation (4.3), COP-RF differs significantly from the conventional information gain function (Breiman, 2001; Liu et al., 2000; Blockeel et al., 1998) as the maximisation of Equation (2.3) is now bounded by the constraint set FL^t. Specifically, the optimisation routine automatically selects discriminative features and their optimal cut-point via feature-information-based energy

optimisation, whilst at the same time fulfilling the guiding conditions imposed by pairwise constraints, leading to semantically adapted data partitions.

More concretely, a data split in COP-tree can be considered as candidate if and only if it respects all involved must-links, i.e. the constrained two samples by some must-link have to be grouped together. Moreover, candidate data splits that fulfil more cannot-links are preferred. The difference in treating must-links and cannot-links originates from their distinct inherent properties: (1) Once a particular must-link is violated at some split node, i.e. the two linked samples are separated apart, there will be no chance to compensate for agreeing again with this must-link in the subsequent process; That means all must-links have to be fulfilled anytime. (2) Whilst a cannot-link would be fulfilled forever once it is respected one time. This property allows us to ignore a cannot-link temporarily. In particular, although the learning process prefers data splits that fulfil more cannot-links, it does not need to forcefully respect all cannot-links at the current split node. Algorithm 1 summarises the split function optimisation procedure in a COP-tree.

Once a COP-RF model is trained, the data affinity matrix can be computed by using ClustRF-Bi as described in Section 3.1.1. ClustRF-Strct-Unfm and ClustRF-Strct-Adpt are not considered, due to (1) ClustRF-Bi has the cheapest computation cost among the three methods, and (2) the essential problem of this study is not on clustering forest based affinity models. This affinity matrix already encodes the knowledge from the available pairwise links since the pairwise constraints have been enforced during the COP-RF learning process.

Discussion Recall that the data partitions in COP-RF are required to agree with the imposed pairwise constraints, which are defined by splitting conditions in Equation (4.3). From Equation (3.7), it is clear that the pairwise similarity matrix induced by COP-RF is determined by the data partitions formed over its leaves. Hence, the pairwise similarity matrix induced by COP-RF indirectly encodes the pairwise constraints defined by oracles. To summarise, we denote the constraint propagation in COP-RF by the process chain below: *pairwise constraints* $\rightarrow$ *steering data partitions in COP-RF* $\rightarrow$ *distorting pairwise similarity measures*. As the data partitioning operation in COP-RF is driven by the optimal split functions that are defined on discovered discriminative features (Equation (2.1)), the corresponding constraint propagation process takes place naturally in discriminative feature subspaces.

Algorithm 1: Split function optimisation in a COP-tree.

Input: At a split node s of a COP-tree t: (1) Training samples S arriving at a split node s; (2) Pairwise constraints: $FL^t = ML^t \cup CL^t$;

Output: (1) The best feature cut-point $\hat{\boldsymbol{\vartheta}}$, and (2) The associated child node partition $\{\hat{L}, \hat{R}\}$;

Optimisation:
Initialise $L = R = \emptyset$ and $\Delta\psi = 0$;
maxCLs $= 0$; /* the max number of respected cannot-links */
for $var \leftarrow 1$ **to** d_{try} **do**
    Select a feature $x_{\text{var}} \in \{1, \ldots, d\}$ randomly;
    for *each possible cut-point of the feature* x_{var} **do**
        Split S into a candidate partition $\{L, R\}$;
        $dec =$ **validate**$(\{L, R\}, \{ML^t, CL^t\}, \text{maxCLs})$;
        if *dec is true* **then**
            Compute information gain $\Delta\tilde{\psi}$ following Equation (4.3);
            if $\Delta\tilde{\psi} > \Delta\psi$ **then**
                Update $\hat{\boldsymbol{\vartheta}}$;
                Update $\Delta\psi = \Delta\tilde{\psi}$, $\tilde{L} = \tilde{L}$, and $\tilde{R} = \tilde{R}$.
            end
        end
        else
            Ignore the current splitting.
        end
    end
end
if *No valid splitting found* **then**
    A leaf is formed.
end
function **validate**$(\{L, R\}, \{ML, CL\}, \text{maxCLs})$
{
  /* Deal with must-links */
  $\forall(\mathbf{x}_i, \mathbf{x}_j) \in ML$,
  if ($\mathbf{x}_i \in L$ and $\mathbf{x}_j \in R$, or vice versa) return false.
  /* Deal with cannot-links */
  Count the number repCLs of respected cannot-links;
  if (repCLs $<$ maxCLs) return false.
  else maxCLs $=$ repCLs.
  Otherwise, return true.
}

4.1.3 Coping with Imperfect Pairwise Relationships

Most existing models (Wagstaff et al., 2001; Kamvar et al., 2003; Lu and Ip, 2010) assume that all the available pairwise constraints are correct. It is not always so in reality, e.g. annotations from crowd-sourcing are likely to contain invalid constraints due to data ambiguity or mistakes by human. The existence of fault constraints can result in error propagation to neighbouring unlabelled points. To overcome this problem, we formulate a numerical method to measure the quality of individual constraints by estimating their inconsistency with the underlying data distribution, so as to facilitate more reliable constraint propagation in COP-RF.

Incorrect pairwise constraints are likely to conflict with the intrinsic data distributions in the feature space. Motivated by this intuition, we propose an approach to estimating constraint inconsistency measure, as described below.

Specifically, we adopt the outlier detection mechanism offered by classification random forest (Breiman, 2001) to measure the inconsistency of a given constraint. First, we establish a set of samples with $Z = \{\mathbf{z}_i\}_{i=1}^{|FL|}$ with class labels $C = \{c_i\}_{i=1}^{|FL|}$, where $|FL|$ represents the total of constraints. Here, a sample $\mathbf{z}$ is defined as

$$\mathbf{z} = \begin{bmatrix} |\mathbf{x}_i - \mathbf{x}_j| \\ \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_j) \end{bmatrix}, \tag{4.4}$$

where $(\mathbf{x}_i, \mathbf{x}_j)$ is a sample pair labelled with either must-link or cannot-link. We assign $\mathbf{z}$ with class $c = 0$ if the associated constraint is cannot-link, and $c = 1$ for must-link. Equation (4.4) considers both relative position and absolute locations of $(\mathbf{x}_i, \mathbf{x}_j)$. This characteristic enables the forest learning process to be position-sensitive and thus achieve data-structure-adaptive transformation (Xiong et al., 2012).

Subsequently, we train a conventional classification random forest using Z and C. This learned forest can then be used to measure the inconsistency of each sample $\mathbf{z}_i$. A sample is deemed inconsistent if it is unique against other samples with the same class label. Formally, based on the affinity $\mathbf{A}_{\text{link}}$ on Z that can be computed with Equation (3.7) and Equation (3.5) using this conventional forest, the inconsistency measure ε of $\mathbf{z}_i$ is defined as

$$\varepsilon(\mathbf{z}_i) = \frac{\rho_i - \bar{\rho}}{\bar{\rho}}, \text{ where} \tag{4.5}$$

$$\bar{\rho} = \text{median}([\rho_1, \ldots, \rho_{|Z^i|}]),$$

$$\rho_i = \frac{1}{\sum_{\mathbf{z}_j \in Z^i} (\mathbf{A}_{\text{link}}(\mathbf{z}_i, \mathbf{z}_j))^2},$$

where Z^i comprises of all samples with the same class label as $\mathbf{z}_i$ in Z. By Equation (4.5), we assign a high inconsistency score to $\mathbf{z}_i$ if it has low similarity to samples with the same class label, and a low inconsistency score otherwise. Finally, the inconsistency measure of each constraint $(\mathbf{x}_i, \mathbf{x}_j) \in FL$ is obtained by simply taking the ε of the corresponding $\mathbf{z}$. An overview of the proposed constraint inconsistency quantification is depicted in Algorithm 2.

Algorithm 2: Quantifying constraint inconsistency.

Input: Pairwise constraints: $(\mathbf{x}_i, \mathbf{x}_j) \in FL = \{ML \cup CL\}$;

Output: Inconsistency scores of individual constraints $(\mathbf{x}_i, \mathbf{x}_j) \in FL$;

1 **Quantifying process**:

2 Generate a new sample set $Z = \{\mathbf{z}_i\}_{i=1}^{|FL|}$ with class labels $C = \{c_i\}_{i=1}^{|FL|}$ from constraints FL (Equation (4.4));

3 Train a conventional classification forest with Z and C;

4 Compute an inconsistency score ε for each $\mathbf{z}$ or constraint (Equation (4.5)).

To remove potentially noisy constraints, we rank all the pairwise constraints based on their inconsistency score in an ascending order. Given the rank list, we keep the top $\beta\%$ of the constraints for COP-RF training. In our study, we set $\beta = 50$ obtained by cross-validation.

After computing the affinity matrix by COP-RF (Equation (3.5)), it can be fed into any pairwise similarity based clustering methods, such as spectral clustering (Ng et al., 2002; Zelnik-manor and Perona, 2004; Von Luxburg, 2007; Xiang and Gong, 2008), affinity propagation (Frey and Dueck, 2007). Since the affinity matrix $\mathbf{A}$ is constraint-aware, these conventional clustering models are automatically transformed to conduct constrained clustering on data. For spectral clustering, we generate as model input a k-nearest neighbour graph from $\mathbf{A}$, a typical local neighbourhood graph in spectral clustering literature (Von Luxburg, 2007). Following (Frey and Dueck, 2007), we perform affinity propagation directly on $\mathbf{A}$. In Section 4.3, we will show extensive experiments to demonstrate the effectiveness of the proposed COP-RF in constrained clustering.

4.1.4 COP-RF Model Complexity Analysis

COP-trees in a COP-RF model can be trained independently in parallel, as in most of the random forest models. For the worst case complexity analysis, here we consider a sequential training mode, i.e. each tree is trained one after another with a 1-core CPU.

The learning complexity of a whole COP-RF can be examined from its constituent parts. Specifically, it can be decomposed into tree- and node-levels as: (i) The complexity of learning a

COP-RF is directly determined by individual COP-tree training costs. (ii) Similarly, the training time of a single COP-tree relies on the costs of learning individual split nodes. Formally, given a COP-tree t, we denote the set of all the internal nodes as Π_t and the sample subset used for training an internal node $s \in \Pi_t$ as S, the training complexity of s is then $d_{\text{try}}(|S|-1)u$ when a greedy search algorithm is adopted, with d_{try} the number of features attempted to partition S during training s, and u the complexity of conducting one data splitting operation. As shown in Algorithm 1, the cost of a single data partition in a COP-tree includes two components: (1) the validation of constraint satisfaction; and (2) the computation of information gain. Therefore, the overall computational cost of learning a COP-RF can be estimated as

$$\Omega = \sum_{t}^{\tau_{\text{clust}}} \sum_{s \in \Pi_t} d_{\text{try}}|S|u = d_{\text{try}}u \sum_{t=1}^{\tau_{\text{clust}}} \Phi(t), \tag{4.6}$$

where

$$\Phi(t) = \sum_{s \in \Pi_t} |S-1| \tag{4.7}$$

is called *tree fan-in*, and τ_{clust} is the number of trees in a COP-RF. Note that the value of $\Phi(t)$ depends on both the training sample size n and the tree topological structure, so it is difficult to express in an explicit form if possible. In Section 4.3.4 we will examine the actual run time needed for training a COP-RF.

4.2 Experimental Settings

Evaluation metrics We used the widely adopted Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) as evaluation metric. ARI measures the agreement between the cluster results and the ground truth in a pairwise fashion, with higher values indicating better clustering quality in the range $[-1,1]$. Throughout all experiments, we reported the ARI values averaged over 10 trials. In each trial we generated a random pairwise constraint set from the ground truth cluster labels.

Implementation details The number of trees, τ_{clust}, in a COP-RF is set to 1000. In general, we found that better results can be achieved by adding more trees, in line with the observation in (Criminisi and Shotton, 2012). Each X^t is obtained by performing n times of random selection with replacement from the augmented data space of $2 \times n$ samples (see Section 2.2). The depth of each COP-tree is governed by either constraint satisfaction, i.e. a node will stop growing if during any attempted data partitioning constraint validation fails (see Algorithm 1), or the size of a node equals to 1 (i.e. $\phi = 1$). We set d_{try} (see Equation (2.2)) to $\sqrt{d}$ with d the feature dimensionality

Table 4.1: Dataset details.

Dataset	# Clusters	# Features	# Instances
ERCe	6	2672	600
Ionosphere (Iono.)	2	34	351
Iris	3	4	150
Segmentation (Seg.)	7	19	210
Parkinsons (Park.)	2	22	195
Glass	6	10	214

of the input data and employ a axis-aligned data separation (Criminisi and Shotton, 2012) as the split function (see Equation (2.1)). More complex split functions, e.g. quadratic functions or Support Vector Machine (SVM), can be adopted at a higher computational cost. We set $k \approx n/10$ (n is the dataset size) for the k-nearest neighbour graph construction in the constrained spectral clustering experiments.

4.3 Experiments and Evaluations

4.3.1 Evaluation on Spectral Clustering

Datasets We utilised an intrinsically noisy video dataset from a publicly available web-camera deployed in a university's Educational Resource Centre (ERCe). This video dataset is challenging as it contains a wide range of physical events characterised by large changes in the environmental set-up, participants, and crowdedness, as well as intricate activity patterns. It also potentially contains large amount of noise in its high-dimensional feature space. The dataset consists of 600 video clips with six possible clusters of events, namely Student Orientation, Cleaning, Career Fair, Gun Forum, Group Studying, and Scholarship Competition (see Figure 3.2c for example images). To evaluate the effectiveness of our method in coping with diverse forms of data with varying numbers of dimensions and clusters, we also selected five UCI benchmark datasets (Asuncion and Newman, 2007), which have been widely employed to evaluate clustering and classification techniques. The details of all datasets are summarised in Table 4.1.

Data feature representation For the UCI datasets, we used the original features provided. As for the ERCe video data, we segmented a long video into non-overlapping clips (each consisting

of 100 frames), from which a number of visual features were then extracted, including colour features (RGB and HSV), local texture features (LBP) (Ojala et al., 2002), optical flow, image features (GIST) (Oliva and Torralba, 2001), and person detections (Felzenszwalb et al., 2010). The resulting 2672-D feature vectors of video clips may contain a large number of less informative dimensions, we performed PCA based linear transform on them and the first 30 PCA components are used as the final feature representation. While it is known that random forest is characterised in feature selection and handling noisy data, this pre-processing step allows to obtain strong components/features by linear projection in addition to removing possibly noisy ones, which is significant for random forests whose convergence rate depends largely on the number of strong features (Biau, 2012). As empirically revealed, without this PCA transform, weaker results by COP-RF can be yielded, possibly due to (1) there may be only a few number of strong dimensions in the original visual features; (2) the dataset is small and thus only allows shallow trees to be grown and in turn not sufficient in exploiting the informative features given a high-dimensional data vector. For all compared methods, we utilised the same feature data, which were linearly scaled to the range of $[-1,1]$.

Baselines For comparison, we presented the results of the baselines[2] as below:

1. *Spectral Clustering (SPClust)* (Ng et al., 2002): the conventional spectral clustering algorithm without exploiting pairwise constraints.
2. *COP-Kmeans* (Wagstaff et al., 2001): a popular constrained clustering method based on *k*-means. The algorithm attempts to satisfy all pairwise constraints during the iterative refinement of clusters.
3. *Spectral Learning (SL)* (Kamvar et al., 2003): a constrained spectral clustering method without constraint propagation. It extends SPClust by trivially adjusting the elements in a data affinity matrix with 1 and 0 to satisfy must-link and cannot-link constraints, respectively.
4. *E^2CP* (Lu and Ip, 2010): a state-of-the-art constrained spectral clustering approach, in which constraint propagation is achieved by manifold diffusion (Zhou et al., 2004). We use the original code released by (Lu and Ip, 2010), with parameter setting as suggested by the paper, i.e. we set the propagation trade-off parameter as 0.8.

[2] We experimented the constrained clustering method in (Coleman et al., 2008) which turns out to produce the worst performance across all datasets, and thus ignored in our comparison.

Table 4.2: Comparing different methods by the area under the curve of ARI score against neighbourhood size. Perfect oracles are assumed. Higher is better.

Method	SPClust	COP-Kmeans	SL	E^2CP	RF+E^2CP	COP-RF
Ionosphere	0.490	0.225	0.063	0.176	**3.120**	2.979
Iris	3.273	1.632	3.499	**3.516**	3.265	3.385
Segmentation	1.943	0.499	1.973	1.989	**2.266**	2.239
Parkinsons	0.677	0.114	0.811	0.787	1.082	**1.403**
Glass	1.121	0.394	1.162	1.210	1.602	**2.015**
ERCe	2.647	0.292	3.681	3.447	3.840	**3.947**
Average	1.692	0.526	1.865	1.854	2.529	**2.661**

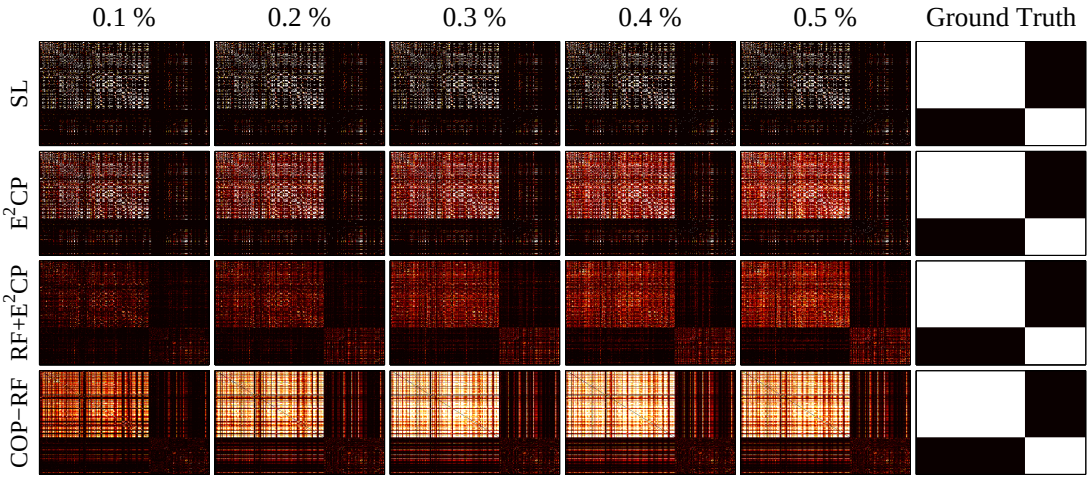
5. *RF+E^2CP*: we modified E^2CP (Lu and Ip, 2010), i.e. instead of generating the data affinity matrix with Euclidean-based measure, we use a conventional clustering forest (equivalent to a COP-RF without constraints imposed and noisy constraint filtering mechanism) to generate the affinity matrix. The constraint propagation is then performed using the original E^2CP-based manifold diffusion. This allows E^2CP to enjoy a limited capability of feature selection using a random forest model.

We carried out comparative experiments to (1) evaluate the effectiveness of different clustering methods in exploiting sparse but perfect pairwise constraints (Section 4.3.1), and (2) compare their clustering performances in the case of having imperfect oracles to provide ill-conditioned pairwise constraints (Section 4.3.1).

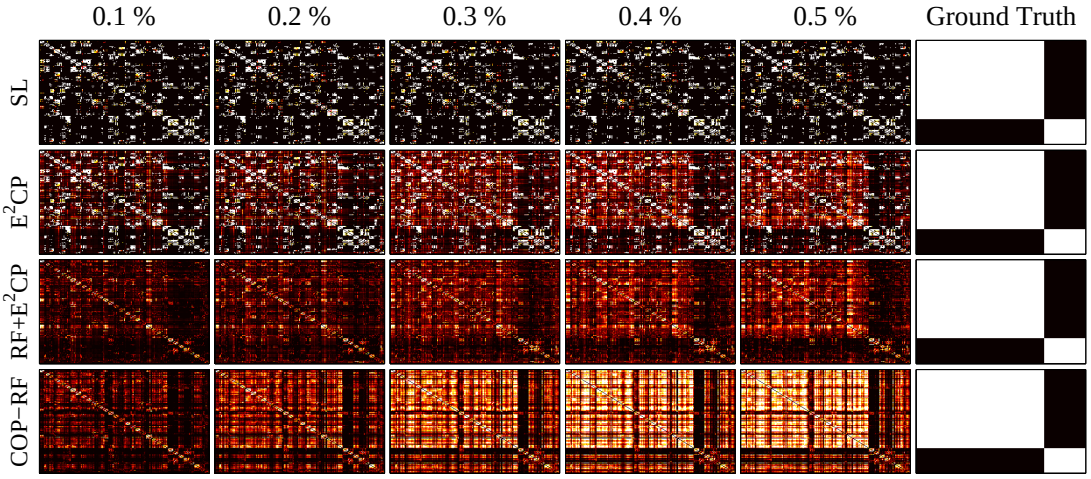
Evaluation on Sparse Constraint Propagation

In this experiment, we assume perfect oracles thus all the pairwise constraints agree with the ground truth cluster labels. First, we examined the data affinity matrix after employing the available constraints, which may reflect how effective a constrained clustering method is. Figure 4.3 depicts some examples of affinity matrices produced by SL, E^2CP, RF+E^2CP, and COP-RF, respectively. COP-Kmeans is excluded since it is not a spectral method. It can be observed that COP-RF produces affinity matrices with more distinct block structure in comparison to its competitors on the most cases. Moreover, the block structure becomes clearer when more pairwise constraints are considered. The results demonstrate the superiority of the proposed approach in propagating sparse pairwise constraints, leading to more compact and separable clusters.

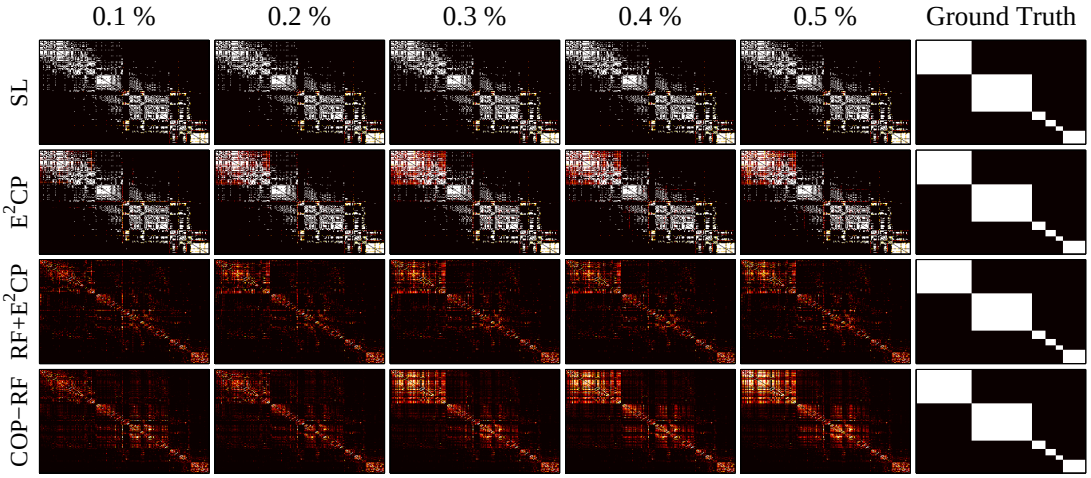
Figure 4.4 reports the curves of ARI score against neighbourhood size by different methods



(a) Ionosphere

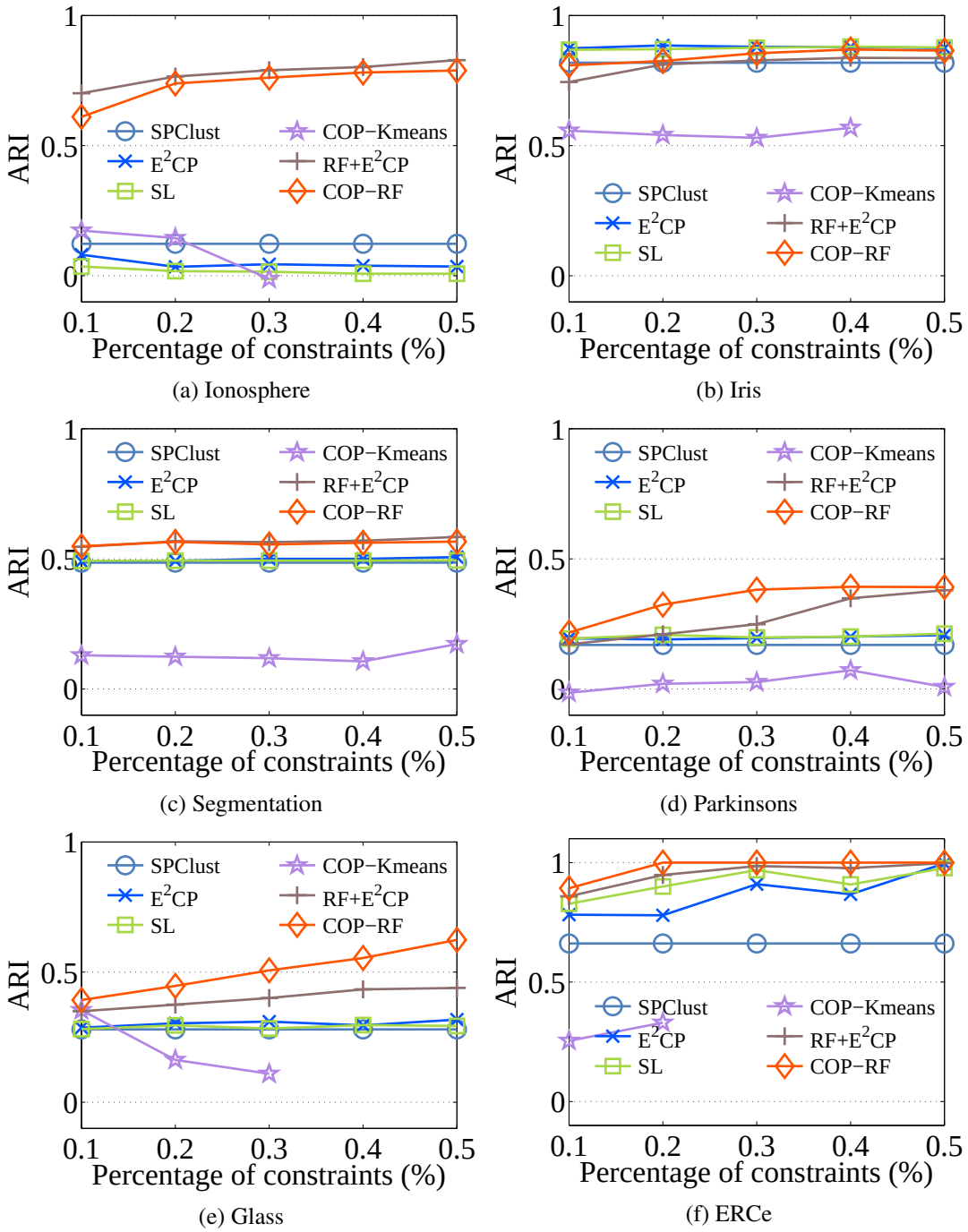


(b) Parkinsons



(c) Glass

Figure 4.3: Comparison of affinity matrices by different methods given a varying number (0.1 $\sim$ 0.5%) of perfect pairwise constraints.



(a) Ionosphere (b) Iris

(c) Segmentation (d) Parkinsons

(e) Glass (f) ERCe

Figure 4.4: ARI score comparison of clustering performance between different methods given a varying number of perfect pairwise constraints.

along with varying numbers of pairwise constraints (ranging in $0.1 \sim 0.5\%$ of total constraints $\frac{n(n-1)}{2}$ where n is the number of data samples). The overall performance of various methods can be quantified by the area under the curves (AUC). This ideal/full AUC area is the integral of best accuracy (i.e. ARI = 1) over all 5 constraint percentages (or 4 durations) on the X-axis, which is 4. The results are reported in Table 4.2. It is evident from the results (Figure 4.4 and Table 4.2) that on most datasets, the proposed COP-RF outperforms other baselines, by as much as $>$**400**% against COP-Kmeans and $>$**40**% against the state-of-the-art E^{2}CP in averaged area under the ARI score curve. This is in line with our previous observations on the affinity matrices (Figure 4.3). Unlike E^{2}CP that relies on the conventional Euclidean-based affinity matrix that considers all features for constraint propagation, COP-RF propagate constraints via discriminative subspaces (Section 4.1.2), leading to its superior clustering results. Additionally, we examined and compared the number of same-cluster sample pairs consistent with the ground truth. Specifically, there are a total of 89088 true same-cluster pairs over all the six datasets, among which COP-RF and RF+E^{2}CP discovered 77637 and 75283, respectively. In other words, the proposed COP-RF found 3.13% relatively more truth pairs than RF+E^{2}CP.

We now examine and discuss the performance of other baselines. The poorest results are given by COP-Kmeans on majority datasets, beyond which some incomplete curves are observed in Figure 4.4 as the model fails to converge (early termination without a solution) as more constraints are introduced into the model. On the contrary, COP-RF is empirically more stable than COP-Kmeans, as COP-RF casts the difficult constraint optimisation task into smaller subproblems to be addressed by individual trees. This characteristic is reflected in Equation (4.2), where each tree in a COP-RF only needs to consider a subset of constraints $FL^t \subset FL$.

SPClust's performance is surprisingly better than COP-Kmeans although it does not utilise any pairwise constraint. This may be because of: (1) in comparison to the conventional k-means, SPClust is less sensitive to noise as it partitions data in a low-dimensional spectral domain (Von Luxburg, 2007), and (2) the limited ability of COP-Kmeans in exploiting pairwise constraints. SL performs slightly better than SPClust through switching the pairwise affinity value in accordance to must-link and cannot-link constraints. Due to the lack of constraint propagation, SL is less effective in exploiting limited supervision information when compared to propagation based models.

Better results are obtained by constraint propagation based E^{2}CP. Nevertheless, the state-of-

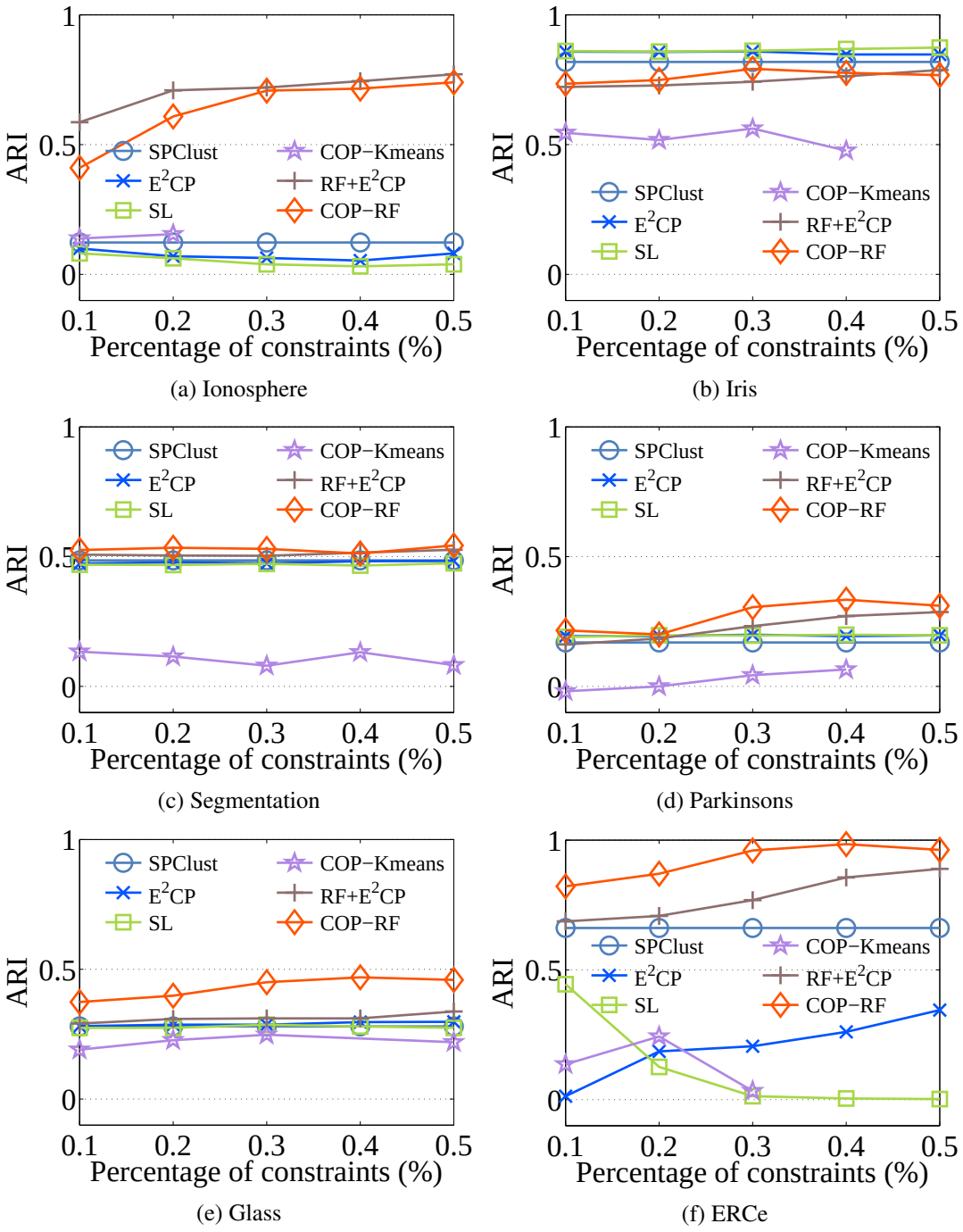
the-art E^{2}CP is inferior to the proposed COP-RF, since its manifold construction still considers the full feature space, which may be corrupted by noisy features. We observe in some cases, such as the challenging ERCe dataset, the performance of E^{2}CP is worse than the naive SL method that comes without constraint propagation. This result suggests that propagation could be *harmful* when the feature space is noisy. The variant modified by us, i.e. RF+E^{2}CP, employs a conventional clustering forest ((Blockeel et al., 1998; Liu et al., 2000)) to generate the data affinity matrix. This allows E^{2}CP to take advantage of a limited capability of forest-based feature selection, and better results are obtained compared with the pure E^{2}CP. Nevertheless, RF+E^{2}CP's performance is generally poorer than COP-RF's (see Table 4.2). This is because the feature selection of the ordinary forest model is less effective than that of COP-RF, which jointly considers feature-based information gain maximisation and constraint satisfaction.

To further highlight the superiority of COP-RF, we show in Figure 4.6 the improvement of area under the ARI score curve achieved by COP-RF relative to other methods (dark bars). Clearly while COP-RF rarely performs noticeably worse than the others, the potential improvement is large.

Evaluation on Propagating Noisy Constraints

In this experiment, we assume imperfect oracles thus pairwise constraints are noisy. We conduct two sets of comparative experiments: (1) We deliberately introduced a fixed ratio (15%) of random invalid constraints into the perfect constraint sets as used in the previous experiment (Section 4.3.1). This is to simulate the annotation behaviour of imperfect oracles for the comparison of our approach with existing models. (2) Given a set of random constraints sized 0.3% of the total constraint samples, we varied the quantity of random noisy constraints, e.g. from 5% to 30%. This allows us to further compare the robustness of different models against mistaken pairwise constraints. In both experiments, we repeated the same experimental protocol as discussed in Section 4.3.1.

A fixed ratio of noisy constraints In this evaluation, we examined the performance of different models when 15% of noisy constraints are included in the given constraint sets. The performance comparison are reported in Figure 4.5 and Table 4.3 and the relative improvement in Figure 4.6. It is observed from Table 4.3 that in spite of the imperfect oracle assumption, COP-RF again achieves better results than other constrained clustering models on most datasets as well as the best average clustering performance across datasets, e.g. $>$**300**% increase against COP-Kmeans



(a) Ionosphere (b) Iris

(c) Segmentation (d) Parkinsons

(e) Glass (f) ERCe

Figure 4.5: ARI score comparison of clustering performance between different methods given a fixed (15%) ratio of invalid constraints.

Table 4.3: Comparing different methods by the area under the ARI score curve against neighbourhood size. A fixed ratio (15%) of invalid pairwise constraints are involved. Higher is better.

Method	SPClust	COP-Kmeans	SL	E^2CP	RF+E^2CP	COP-RF
Ionosphere	0.490	0.146	0.192	0.276	**2.851**	2.606
Iris	3.273	1.590	**3.454**	3.416	2.988	3.067
Segmentation	1.943	0.433	1.877	1.913	2.039	**2.109**
Parkinsons	0.677	0.067	0.786	0.780	0.910	**1.102**
Glass	1.121	0.679	1.114	1.159	1.244	**1.734**
ERCe	2.647	0.328	0.368	0.832	3.119	**3.705**
Average	1.692	0.540	1.299	1.396	2.192	**2.387**

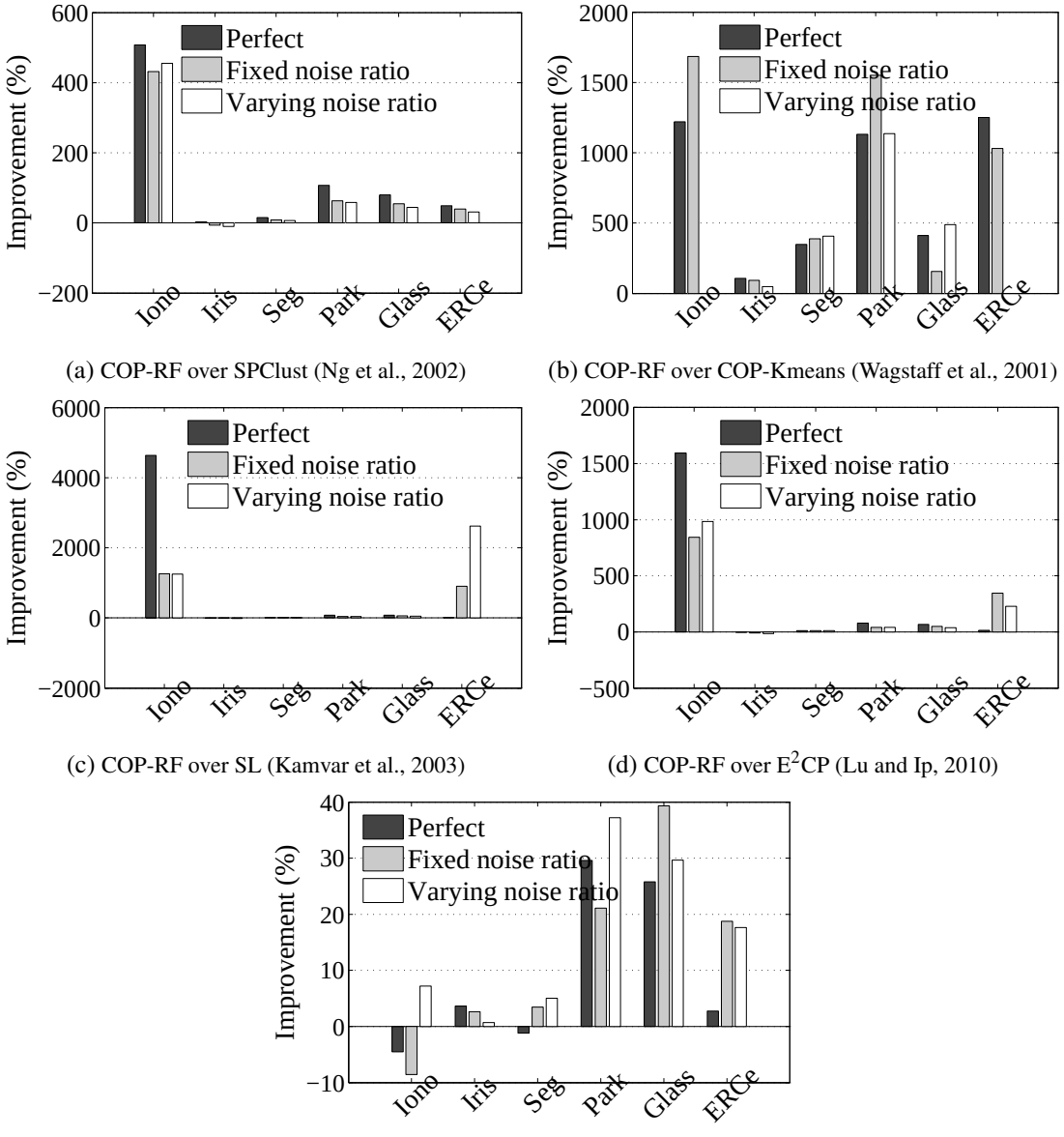
Table 4.4: Evaluating the generic effect of the proposed constraint inconsistency measure by AUC. A fixed ratio (15%) of invalid pairwise constraints are involved. Higher is better.

Dataset	Ionosphere	Iris	Segmentation	Parkinsons	Glass	ERCe	*Average*
RF+E^2CP	**2.851**	2.988	2.039	0.910	1.244	3.119	2.192
RF+E^2CP (Filtered)	2.568	3.003	2.086	0.915	1.350	3.508	2.238
COP-RF	2.606	**3.067**	**2.109**	**1.102**	**1.734**	**3.705**	**2.387**

and $>$**70**% increase against E^2CP. Furthermore, Figure 4.6 also shows that COP-RF maintains encouraging performance given noisy constraints, in some cases such as the challenging ERCe video dataset even larger improvements are obtained over E^2CP and other models, compared with the perfect constraint case.

We further evaluated the generic effect of our constraint inconsistency measure algorithm for baseline methods, e.g. RF+E^2CP. The results are presented in Table 4.4. It is observed that by filtering potentially constraints, RF+E^2CP is able to produce more accurate clustering results in most cases, with the specific improvement differs over datasets. This suggests the general usefulness of our constraint inconsistency measure. Overall, the proposed COP-RF still achieves the best average performance.

Varying ratios of noisy constraints Noisy constraints bring negative impact on the clustering results, as shown in the above experiment. We wish to investigate how constrained clustering models would perform under different ratios of noisy constraints. To this end, we evaluated the robustness of compared models against different amounts of noisy constraints involved in sets of 0.3% out of the full pairwise constraints. Figure 4.7 and Table 4.5 show that COP-RF once again



(a) COP-RF over SPClust (Ng et al., 2002)

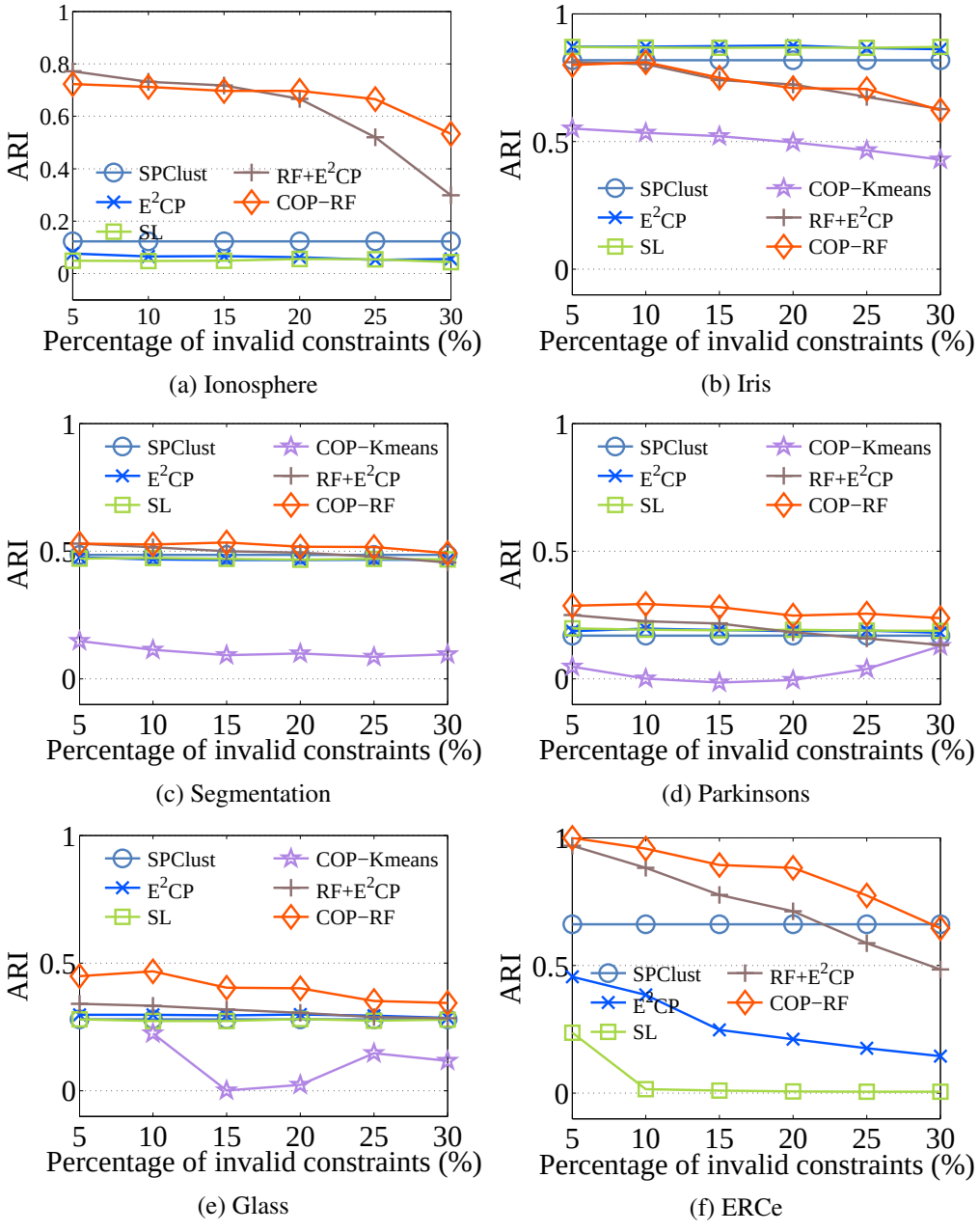
(b) COP-RF over COP-Kmeans (Wagstaff et al., 2001)

(c) COP-RF over SL (Kamvar et al., 2003)

(d) COP-RF over E^2CP (Lu and Ip, 2010)

(e) COP-RF over RF+E^2CP

Figure 4.6: The improvement of AUC achieved by COP-RF relative to baseline methods. Dark bars: when perfect constraints are provided. Grey bars: when 15% of the total pairwise constraint samples are noisy. White bars: when varying ratios ($5 \sim 30\%$) of noisy constraints are provided.



(a) Ionosphere (b) Iris

(c) Segmentation (d) Parkinsons

(e) Glass (f) ERCe

Figure 4.7: ARI score comparison of clustering performance between different constraint propagation methods given varying ratios of invalid constraints.

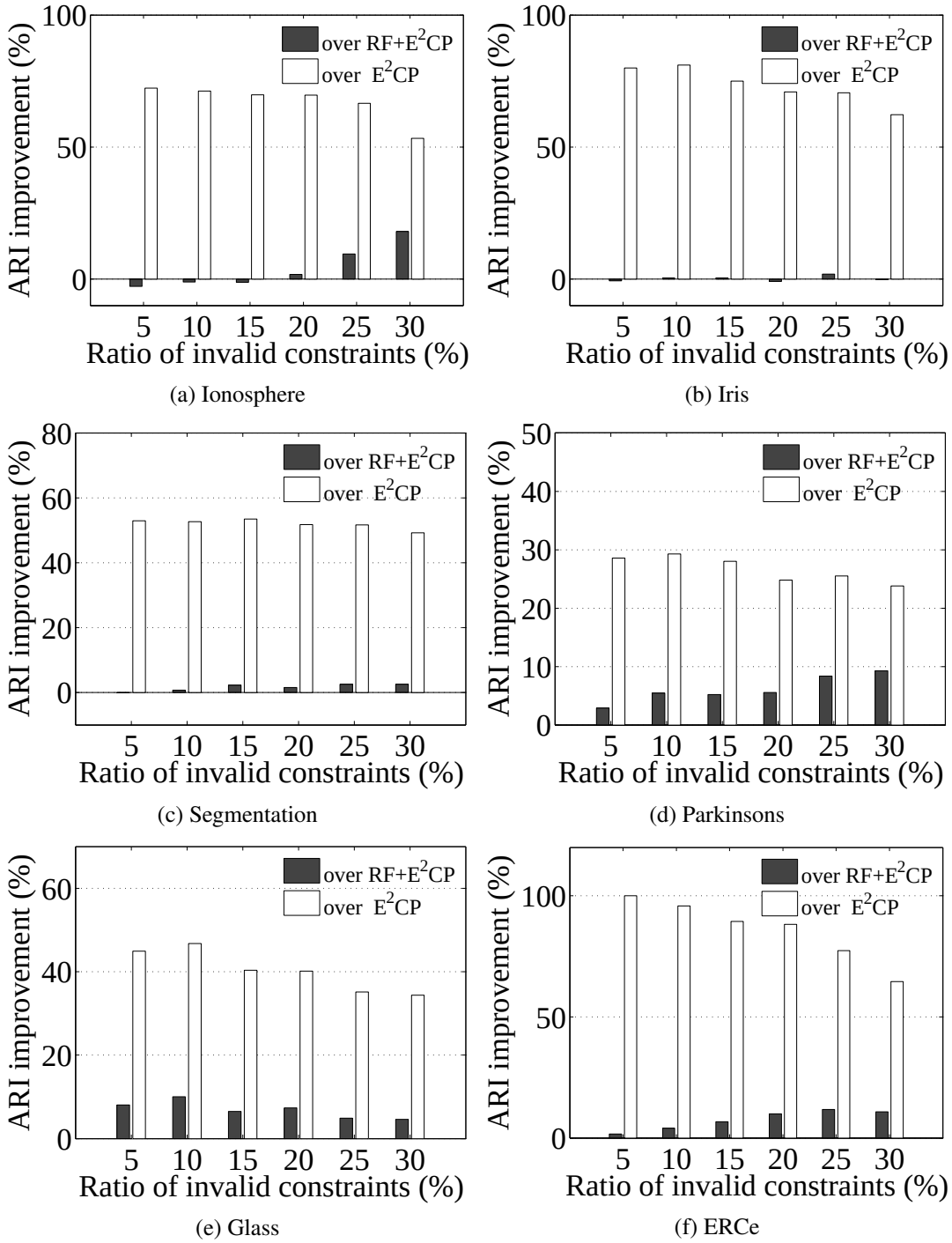


Figure 4.8: ARI relative improvement of COP-RF over baseline constraint propagation models given varying ratios of noisy constraints in 0.3% out of the full constraints. Higher is better.

Table 4.5: Comparing different methods by the area under the ARI score curve against neighbourhood size. Varying ratios ($5 \sim 30\%$) of invalid pairwise constraints are involved. Higher is better.

Method	SPClust	COP-Kmeans	SL	E^2CP	RF+E^2CP	COP-RF
Ionosphere	0.536	0.000	0.253	0.314	3.172	**3.399**
Iris	4.341	2.507	4.339	**4.352**	3.659	3.684
Segmentation	2.462	0.514	2.348	2.336	2.481	**2.605**
Parkinsons	0.979	0.108	0.957	0.948	0.975	**1.338**
Glass	1.421	0.343	1.380	1.477	1.558	**2.020**
ERCe	3.160	0.000	0.159	1.320	3.682	**4.331**
Average	2.150	0.579	1.573	1.791	2.588	**2.896**

outperforms the competitor models on most datasets. As shown in Figure 4.8, the performance improvement of COP-RF over constraint propagation baselines maintains over varying degrees of noisy constraints in most cases. Specifically, COP-RF's average relative improvements over E^2CP and RF+E^2CP across all datasets are **63**% and **2**% given 5% noisy constraints whilst **48**% and **8**% given 30% noise.

4.3.2 Evaluation on Affinity Propagation

To demonstrate the generalisation of our COP-RF model, we showed its effectiveness on affinity propagation, an exemplar-location based clustering algorithm (Frey and Dueck, 2007). Similarly, ARI was used as performance evaluation metrics[3].

Dataset We selected the same face image set as (Frey and Dueck, 2007), which was extracted from the Olivetti database. Particularly, this dataset includes a total of 900 grey images with resolution of 50×50 from 10 different persons, each with 90 images obtained by Gaussian smoothing and rotation/scaling transformation. It is challenging to distinguish these faces (Figure 4.9) due to large variations in lighting, pose, expression and facial details (glasses / no glasses). The features of each image are normalised pixel values with mean 0 and variance 0.1.

Baselines Typically, negative squared Euclidean distance is used to measure the data similarity. Here, we compared the COP-RF model against

[3] Average Squared Error (ASE) is adopted in (Frey and Dueck, 2007) as evaluation metric. This metric requires all comparative methods to produce affinity matrices based on a particular type of similarity/distance function. In our experiments ASE is not applicable since distinct affinity matrices are generated by different comparative methods.

Figure 4.9: Example face images from 10 different identities. Two distinct individuals are included in each row, each with 10 face images.

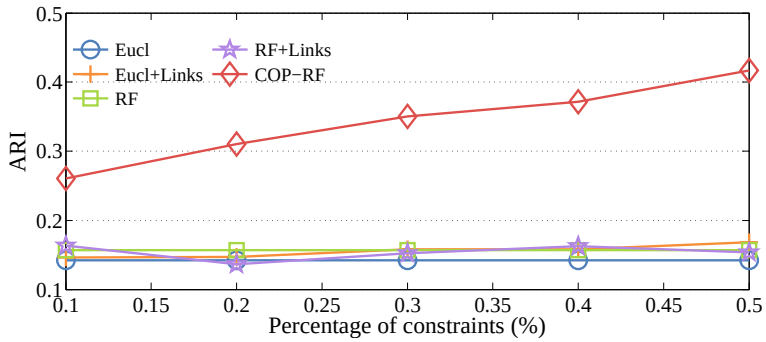


Figure 4.10: Comparison of different methods on clustering faces with affinity propagation.

1. *Eucl*: the Euclidean metric;
2. *Eucl+Links*: we encoded the information of pairwise constraints into the Euclidean-metric based affinity matrix by making the similarity between cannot-linked pairs be minimal and the similarity between must-linked pairs be maximal, similar to (Kamvar et al., 2003);
3. *RF*: the conventional clustering Random Forest (Breiman, 2001) so that the pairwise similarity measures can benefit from feature selection;
4. *RF+Links*: analogues to Eucl+Links but with the affinity matrix generated by the clustering forest.

In this experiment, we used the perfect pairwise links ($0.1 \sim 0.5\%$) as constraints, similar to Section 4.3.1. The results are reported in Figure 4.10. It is evident that the feature selection based similarity (i.e. RF) is favourable over the Euclidean metric that considers the whole feature spaces. This observation is consistent with the earlier findings in Section 4.3.1. Manipulating affinity matrix naively using sparse constraints helps little in performance, primarily due to the lack of constraint propagation. The superiority of COP-RF over all the baselines justifies the effectiveness of the proposed constraint propagation model in exploiting constraints for facilitating cluster formation. Also, obviously larger performance margins are acquired when one increases

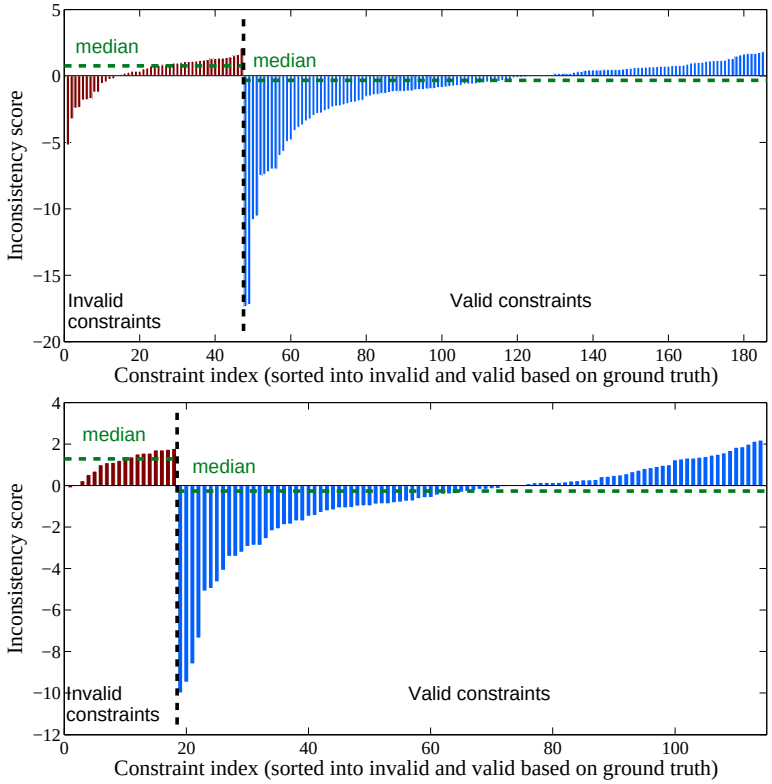


Figure 4.11: Quantifying constraint inconsistency by using the proposed algorithm on Ionosphere (top) and Glass (bottom). See details in Section 4.1.3. High values suggest large probabilities of being invalid constraints.

the amount of pairwise constraints, further suggesting the effectiveness of constraint propagation by the proposed COP-RF model.

4.3.3 Evaluation on Constraint Inconsistency Measure

The superior performance of COP-RF in handling imperfect oracles can be better explained by examining more closely the capability of our constraint inconsistency quantification algorithm (Equation (4.5)). Figure 4.11 shows the inconsistency measures of individual pairwise constraints on Ionosphere and Glass datasets. It is evident that the median inconsistency scores induced by invalid/noisy constraints are much higher than that by valid ones.

4.3.4 Computational Cost Analysis

In this section, we reported the computational complexity of our COP-RF model. Time was measured on a Linux machine of Intel Quad-Core CPU @ 3.30GHz and 8.0GB with C++ implementation of COP-RF. Note that only one core was utilised during the model training procedure. Time analysis was conducted on the ERCe dataset using the same experimental setting as stated

in Section 4.3.1. A total of 60 repetitions were performed, each utilising 0.3% out of the full constraints with varying ($5\% \sim 30\%$) amounts of invalid ones. On average, training a COP-RF takes 213 seconds. Note that the above process can be conducted in parallel in a cluster of machines to speed up the model training.

4.4 Summary

This chapter has presented a semi-supervised visual data structure discovery or constrained clustering framework to (1) propagate sparse pairwise constraints effectively, and (2) handle noisy constraints generated by imperfect oracles. There has been little work that considers these two closely-related challenging problems jointly. The proposed COP-RF model is novel in that it propagates constraints more effectively via discriminative feature subspaces. This is in contrast to existing methods that perform propagation considering the whole feature space, which may be misled by noisy features. Effective propagation regardless of the constraint quality could lead to poor clustering results. Our work addresses this crucial issue by formulating a statistical algorithm to quantify the inconsistency of constraints and effectively perform selective constraint propagation. The model is flexible in that it generates a constraint-aware affinity matrix that can be used by the existing pairwise similarity measure based clustering methods for readily performing constrained data clustering, e.g. spectral clustering, affinity propagation. Experimental results demonstrated the general effectiveness and advantages of the proposed structure analysis approach over the state-of-the-art methods on both visual and non-visual datasets.

The cluster structure analysis models presented in Chapters 3 and 4 are limited to effectively coping with a single source, e.g. visual data alone, at a time. However, other data sources (e.g. traffic condition, weather) may be readily available in the real surveillance scenarios, together with the widely exploited visual source. Considering such multi-source data simultaneously and learning them jointly in a unified way may bring additional benefits to visual data structure discovery, e.g. because these sources are mutually complementary, even through additional challenges may arise due to the large disparity between sources. The next chapter solves a multi-source data structure discovery problem by developing a principled video cluster analysis method capable of combining and learning effectively heterogeneous data sources. The strength of that multi-source video clustering approach is validated in semantically summarising surveillance video streams, a critical video surveillance application.

Chapter 5

Multi-Source Data Structure Discovery for Video Summarisation

---

The previous two chapters describe visual data structure analysis and discovery by unsupervised and semi-supervised clustering. The target data considered by both models are assumed to be drawn from a single data source, e.g. surveillance camera. On the other hand, there may exist a number of auxiliary non-visual data sources that provide complementary perceptions to visual source in video surveillance. As discussed in Section 1.2, the latent correlation between visual and non-visual data may help video data structure discovery in spite of the intrinsic challenges of jointly learning these heterogeneous data with great difference in representation and statistics.

This chapter presents a multi-source data structure discovery framework capable of performing joint learning over multiple heterogeneous visual and non-visual source data (Figure 5.1). Specifically, this approach seamlessly uncovers latent correlations among heterogeneous types of sources, despite the non-trivial heteroscedasticity and dimensionality discrepancy problems. Additionally, the proposed model is robust to partial or missing non-visual information. The effectiveness of this video structure analysis method was demonstrated in performing semantic video summarisation on two crowded public surveillance datasets.

The organisation of this chapter is as follows. Section 5.1 explains the details of the proposed multi-source data clustering model The following is video summarisation (Section 5.2). After describing the datasets and experimental settings (Section 5.3), experiments and evaluations of the proposed model are provided in Section 5.4. A summary is given in Section 5.5.

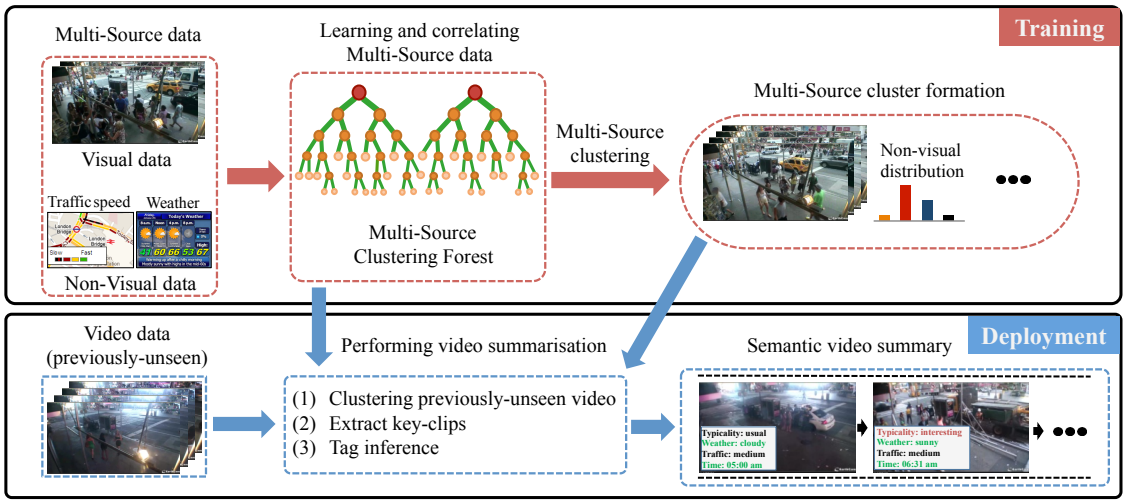


Figure 5.1: Overview of the proposed multi-source driven video structure discovery method and its application on video summarisation. We consider a novel setting where multiple heterogeneous sources are present during the model training stage. The proposed Multi-Source Clustering Forest discovers and exploits latent correlations among heterogeneous visual and non-visual data sources both of which can be inaccurate and not trustworthy. In deployment, our model uncovers visual content structures and infer tags on previously-unseen video data for video summarisation.

5.1 Multi-Source Data Structure Discovery

Video summarisation by content abstraction aims to generate a compact summary composed of key/interesting content from a long previously-unseen video for achieving efficient holistic understanding (Truong and Venkatesh, 2007). A common way to establish a video summary is by extracting and then combining a set of key frames or shots. These key contents are usually discovered and selected from clusters of video frames or clips (Truong and Venkatesh, 2007).

In this study, we follow the aforementioned approach but consider not only visual content of video, but also a large corpus of non-visual data collected from heterogeneous independent sources (Figure 5.2(a)). Specifically, through learning latent structure of multi-source data (Figure 5.2(b-c)), we wish to make reference to and/or impose non-visual semantics directly into video clustering without any human manual annotation of video data (Figure 5.2(d)). Formally, we consider the following different data sources that form a multi-source input feature space:

Visual features We segment a training video into n either overlapping or non-overlapping clips, each of which has a duration of $t_{\texttt{clip}}$ seconds. We then extract a d-dimensional visual descriptor from the ith video clip denoted by $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d, i = 1, \ldots, n$.

Non-visual data Non-visual data are collected from heterogeneous independent sources. We collectively represent m types of non-visual data associated with the i-th clip as $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m}) \in$

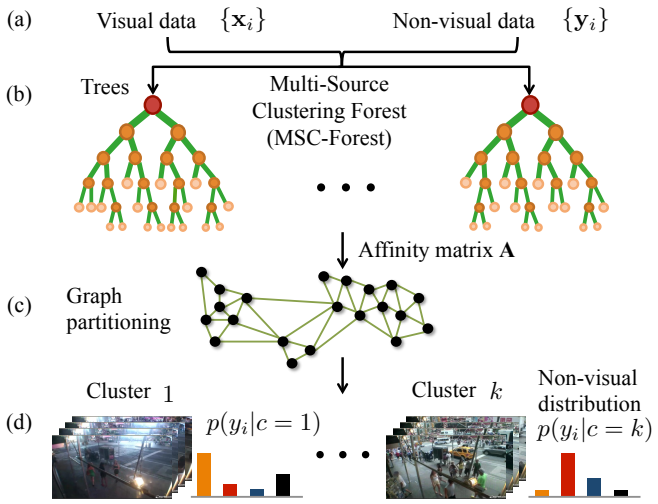


Figure 5.2: Multi-source model training stage: The pipeline of performing multi-source clustering on visual and non-visual data with the proposed Multi-Source Clustering Forest (MSC-Forest).

$\mathbb{R}^m$, $i = 1, \ldots, n$. Note that any (or all) dimension of $\mathbf{y}_i$ may be missing.

We aim at formulating a unified clustering model capable of coping with the few challenges (see Section 1.2). The model needs be unsupervised since no ground truth is assumed. To mitigate the heteroscedasticity and dimension discrepancy problems, we require a model that can isolate the very different characteristics of visual and non-visual data, yet can still exploit their latent correlation in the clustering process. To handle noisy data, feature selection is necessary.

In light of the above demands, we choose to start with the clustering random forest (Breiman, 2001; Liu et al., 2000; Shi and Horvath, 2006) due to (1) unsupervised information gain optimisation thus requiring no ground truth labels; (2) its flexible objective function for facilitating the modelling of multi-source data as well as the processing of missing data; (3) and its implicit feature selection mechanism for handling noisy features. Nevertheless, the conventional clustering forest is not well suited to solve these challenges since it expects a full concatenated representation as input during both model training and deployment. This does not conform to the assumption of only visual data being available during model deployment for previously-unseen videos. Moreover, due to its uniform variable selection mechanism (Breiman, 2001) (e.g. each feature dimension has the same probability to be selected as a candidate optimal splitting variable), there is no principled way to ensure balanced contribution from individual visual and non-visual sources in the node splitting process. To overcome these limitations, we propose a *Multi-Source Clustering Forest* (MSC-Forest) model by introducing an objective function allowing *joint optimisation of individual information gains* of different sources.

5.1.1 Multi-Source Clustering Forest

Conventional clustering forests assumes only homogeneous data sources such as pure imagery-based features. In contrast, the proposed Multi-Source Clustering Forest can take heterogeneous sources as input. In particular, the proposed model uses visual features as splitting variables to grow Multi-Source Clustering trees (MSC-trees) as in Equation (2.1), and exploits non-visual information as additional data to help determining the $\boldsymbol{\vartheta} = [\vartheta_1, \vartheta_2]$. In this way, auxiliary non-visual information is used, in addition to visual data, to guide the tree formation.

Formally, we define a joint information gain function for node splitting during training MSC-trees as:

$$\Delta\psi = \underbrace{\alpha_v \frac{\Delta\psi_v}{\psi_{v0}}}_{\text{visual}} + \underbrace{\sum_{j=1}^{m} \alpha_j \frac{\Delta\psi_j}{\psi_{j0}}}_{\text{non-visual}} + \underbrace{\alpha_t \frac{\Delta\psi_t}{\psi_{t0}}}_{\text{temporal}}. \tag{5.1}$$

Similar to Equation (2.3), the optimal parameter corresponds to the split with the maximal $\Delta\psi$. This formulation defines the best data split across the joint space of multi-source data, beyond visual domain alone. All the terms in Equation (5.1) are interpreted as below.

Visual term $\Delta\psi_v = \Delta\psi_{\text{class}}$ (Equation (2.3)) denotes the information gain in visual domain. Precisely, this measure is computed from the pseudo class labels. Therefore, it reflects the visual data structure characteristics given that the pseudo data samples are drawn from the marginal feature distributions (Section 2.2). In this study we utilise the Gini impurity ψ_{gini} (Breiman et al., 1984) to estimate $\Delta\psi_{\text{class}}$ by setting $\psi = \psi_{\text{gini}}$ in Equation (2.3) due to its simplicity and efficiency. High value in ψ_{gini} (Equation (2.4)) indicates pure category distribution.

Non-visual term This is a new term we introduce as auxiliary information on visual term. More specifically, $\Delta\psi_j$ denotes the information gain in the j-th non-visual data. A non-visual source can be either categorical or continuous. For a categorical non-visual source, similar to visual term we use the Gini impurity ψ_{gini} (Equation (2.4)) as its data split measure criterion. In the case of non-visual source with continuous values, we adopt least squares regression ψ_{lsr} (Equation (2.6)) to enforce continuity in the clustering space,

Temporal term We add a temporal smoothness gain $\Delta\psi_t$ (measured with ψ_{lsr} (Equation (2.6))) to encourage temporally adjacent video clips to be grouped together. This temporal information helps in mining visual data structure.

The information gain by different sources may live in very disparate ranges due to the different natures of source, each term of Equation (5.1) is therefore normalised by its initial data impu-

rity denoted by ψ_{v0}, ψ_{j0}, and ψ_{t0}. These impurities are obtained at the root node of every MSC-tree. The source weights are denoted by α_v, α_i, and α_t accordingly, holding $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$. We set $\alpha_v = 0.5$ obtained by cross-validation. A detailed analysis on α_v is given in Section 5.4.2. For non-visual and temporal information, we uniformly assign $\alpha_t = \alpha_i = \frac{1-\alpha_v}{m+1}$ since their importance is not known *in prior*, with m the number of non-visual sources.

Role of different source data Given the main role and much more stable provision of the visual source in video understanding, non-visual data are regarded as auxiliary information over visual source. During the training of MSC-Forest, the split functions (Equation (2.1)) are defined on visual features, but $\boldsymbol{\vartheta} = [\vartheta_1, \vartheta_2]$ is collectively determined by visual features and the associated non-visual as well as temporal information (i.e. the non-visual and temporal term in Equation (5.1)). Alternatively, one can think of that the *main* visual data source is 'completely-visible' to the MSC-Forest since it is needed during both forest training and evaluation, whilst the *auxiliary* non-visual data are 'half-visible' in that they are exploited as side information for embedding their knowledge into the MSC-tree growing during model training but not required any more during the MSC-Forest evaluation (due to their restricted availability as explained in Section 1.2).

Joint information gain We interpret the intrinsic advantage of the joint information gain defined by Equation (5.1), with comparison against the naïve feature concatenation strategy. With the latter scheme, the information gain (Equation (2.3)) is directly estimated in a heterogeneous joint space where visual, non-visual and temporal data are mixed together. This would suffer from the heteroscedasticity problem, as discussed in Section 1.2. Instead, Equation (5.1) overcomes this challenge by modelling different sources via separate information gain terms, resulting in a more balanced exploitation of multi-source data. In this way, the proposed joint information gain of multi-source data encourages more appropriate visual data separation both visually and semantically. This formulation is the essential contribution of our proposed MSC-Forest model.

Merits of MSC-Forest The formulation in Equation (5.1) brings two unique benefits: (A) Thanks to the information gain optimisation, the influences of visual and non-visual domains on data partitioning can be better balanced compared to naïve feature concatenation. (B) Equation (2.2) and Equation (5.1) together provide a mechanism to discover strongly correlated heterogeneous source pairs and to exploit joint information gain of such correlated pairs for data partitioning. In

other words, only selective visual features (Equation (2.2)) that yield high information gain collectively with non-visual information (Equation (5.1)) will contribute to the MSC-tree growing. Such a mechanism cannot be realised using the conventional clustering forests (Breiman, 2001; Liu et al., 2000). We shall demonstrate the multi-source correlation discovered by our proposed MSC-Forest in experiments (Section 5.4.4).

Coping with Partial/Missing Non-Visual Data

We introduce an adaptive weighting mechanism to dynamically deal with the inevitable partial/missing non-visual data[1]. Specifically, when some non-visual data are missing and suppose the missing proportion of the i-th non-visual type in the training set X_t for MSC-tree t is δ_i, we reduce its weight from α_i to $\alpha_i - \delta_i\alpha_i$. The total reduced weight $\sum_i \delta_i\alpha_i$ is then distributed evenly to the weights of all sources to ensure $\alpha_v + \sum_{i=1}^{m} \alpha_i + \alpha_t = 1$. This linear adaptive weighting method produces satisfactory results in our experiments.

MSC-Forest Model Complexity Analysis

The upper-bound learning complexity of a whole MSC-Forest can be examined in a similar way as the COP-RF model (Section 4.1.4). Clearly, the cost u for conducting a data split in a MSC-Forest is larger than that of conventional forests since we need to compute additional information gains of non-visual and temporal information (Equation (5.1)). On the other hand, the value of $\Phi(t)$ (Equation (4.7)) primarily relies on the tree structure/topological characteristics (Martin, 1997): a balanced and shallower tree has smaller $\Phi(t)$, thus the tree shall be more efficient in training and inference on previously-unseen samples, in that the paths from the root to leaf nodes are relatively shorter. In Section 5.4.5, we will show that the additional non-visual information encourages more balanced and shallower decision trees than learning from single visual source alone.

5.1.2 Latent Multi-Source Data Structure Discovery

Given heterogeneous feature spaces involving visual and non-visual data, it is non-trivial to discover their underlying group structures, due to the heteroscedasticity problem aforementioned (Section 1.2). To this end, MSC-Forest is particularly designed to principally extract and com-

[1] There exist missing data filling algorithms utilised in conventional random forests, e.g. for the missing value of one feature in one class, the median value (continuous) or the most frequent category (discrete) of this feature over the current class can be used as the estimation (Breiman, 2003). Whilst a similar strategy is possible to apply on our MSC-Forest, we consider an alternative by proposing an effective adaptive weighting algorithm in order not to further introduce noisy training data.

bine the information from multiple individual sources so as to more accurately measure data pairwise similarity relations, which in turn facilitates existing graph-based clustering algorithm, e.g. spectral clustering, to eventually reveal the latent data clusters. Figure 5.2 depicts the pipeline of our video data clustering approach based on the learned MSC-Forest.

The spectral clustering (Zelnik-manor and Perona, 2004) groups data using eigenvectors of an affinity matrix derived from the data. The goodness of the resulting cluster formation primarily relies on the quality of the input affinity matrix which reflects and embeds the essential data structures (Zhu et al., 2014). With ClustRF-Bi (Section 3.1.1), we can induce the data affinity matrix from a learned MSC-Forest. Intuitively, the multi-source learning nature of MSC-Forest renders its data similarity measure sensitive to the joint knowledge from diverse source data.

Subsequently, we symmetrically normalise $\mathbf{A}$ to obtain $\mathbf{A}_{\text{norm}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ denotes a diagonal degree matrix with elements $\mathbf{D}_{i,i} = \sum_{j=1}^{n} \mathbf{A}_{i,j}$. Given $\mathbf{A}_{\text{norm}}$, we perform spectral clustering to discover the latent clusters of training clips with the number of clusters automatically determined through analysing the eigenvector structure (Zelnik-manor and Perona, 2004). Each training clip $\mathbf{x}_i$ is then assigned to a cluster $c_i \in C$, with C the cluster index set.

The learned clusters group similar clips both visually and semantically, with each of the clusters associated with a unique distribution for each non-visual data (Figure 5.2(d)). We denote the distribution of the ith non-visual data type of the cluster c as

$$p(y_i|c) \propto \sum_{\mathbf{x}_j \in X_c} p(y_i|\mathbf{x}_j), \tag{5.2}$$

where X_c represents the set of training samples in c. These multi-source data clusters form a component of our multi-source model (Figure 5.1).

5.1.3 Quantifying Correlation between Sources

Quantifying latent correlation between different sources gives insights into their interactions in forming coherent video groupings. This can be done once a MSC-Forest is trained, e.g. all trees have been grown, with the split function in every split node fixed. The primary purpose is to illustrate what correlations between different data sources have been discovered and learned by our MSC-Forest during the training stage, rather than the model learning or optimisation process. To quantify between-source correlation, we first estimate correlation among their constituent features.

Visual-visual feature correlation Visual-visual feature correlation is typically quantified based on their similarity in inducing split node partitions L and R (Breiman, 2001). In particular, given a split node s and its final optimal split, say $L_{\hat{v}}$ and $R_{\hat{v}}$ by a visual feature $\hat{v}$. From Equation (2.2), we recall that this feature $\hat{v}$ is selected out from d_{try} randomly sampled features $\{f_1,\ldots,f_{d_{\text{try}}}\}$. Let $v \in \{f_1,\ldots,f_{d_{\text{try}}}\} \setminus \hat{v}$ and its optimal left-right partitions be L_v and R_v respectively. The node-level correlation between features $\hat{v}$ and v is then defined as

$$\lambda_s(\hat{v},v) = \frac{p_{\hat{v}} - (1 - \frac{|L_{\hat{v}} \cap L_v|}{|L_{\hat{v}} \cup R_{\hat{v}}|} - \frac{|R_{\hat{v}} \cap R_v|}{|L_{\hat{v}} \cup R_{\hat{v}}|})}{p_v}, \tag{5.3}$$

where $p_{\hat{v}} = \min(\frac{|L_{\hat{v}}|}{|L_{\hat{v}}|+|R_{\hat{v}}|}, \frac{|R_{\hat{v}}|}{|L_{\hat{v}}|+|R_{\hat{v}}|})$, thus $p_{\hat{v}} \in (0, \frac{1}{2}]$. With Equation (5.3) we assign a strong correlation ($\lambda_s(\hat{v},v) = 1$) to a feature pair ($\hat{v}$, v) if they produce the same data partition, whilst a weak correlation ($\lambda_s(\hat{v},v) \leq -1$) when their partitions have no overlaps. For simplicity we let $\lambda_s(\hat{v},v) = \max(\lambda_s(\hat{v},v),0)$ such that $\lambda_s(\hat{v},v)$ lies in the range of $[0,1]$. The final visual-visual feature correlation $\lambda(\hat{v},v)$ is obtained via

$$\lambda(\hat{v},v) = \frac{1}{\tau_{\text{clust}}} \sum_{t=1}^{\tau_{\text{clust}}} \left[\frac{1}{n^t_{(\hat{v},v)}} \sum_{k}^{n^t_{(\hat{v},v)}} \lambda_s(\hat{v},v) \right], \tag{5.4}$$

where $n^t_{(\hat{v},v)}$ refers to the number of sampling co-occurrences of a feature pair ($\hat{v}$, v) during the splitting process of a MSC-tree t.

Visual-nonvisual feature correlation Recall that visual and non-visual data play different roles in our MSC-Forest, e.g. the former as splitting features whereas the later as auxiliary information. This difference makes the above equations not applicable to the computation of visual-nonvisual feature correlation since no data split is associated with non-visual features. Instead, we adopt information gain as the visual-nonvisual feature correlation metric. This metric is appropriate in that it also reflects the intrinsic mutual interaction between visual and non-visual features during joint information gain optimisation (Equation (5.1)). Formally, we quantify the node-level correlation between the optimal splitting visual feature $\hat{v}$ and a non-visual feature ω as $\lambda_s(\hat{v},\omega) = \frac{\Delta\psi_\omega}{\psi_{\omega 0}}$ (the non-visual term of Equation (5.1)). The final visual-nonvisual feature correlation $\lambda(\hat{v},\omega)$ is computed similarly by Equation (5.4).

Correlation between sources Given between-feature correlation, the final correlation between any two sources S_i^{src} and S_j^{src} can then be estimated through

$$\varphi(S_i^{\text{src}}, S_j^{\text{src}}) = \frac{1}{|S_i^{\text{src}}||S_j^{\text{src}}|} \sum_{v_i \in S_i^{\text{src}}, v_j \in S_j^{\text{src}}} \lambda(v_i, v_j). \tag{5.5}$$

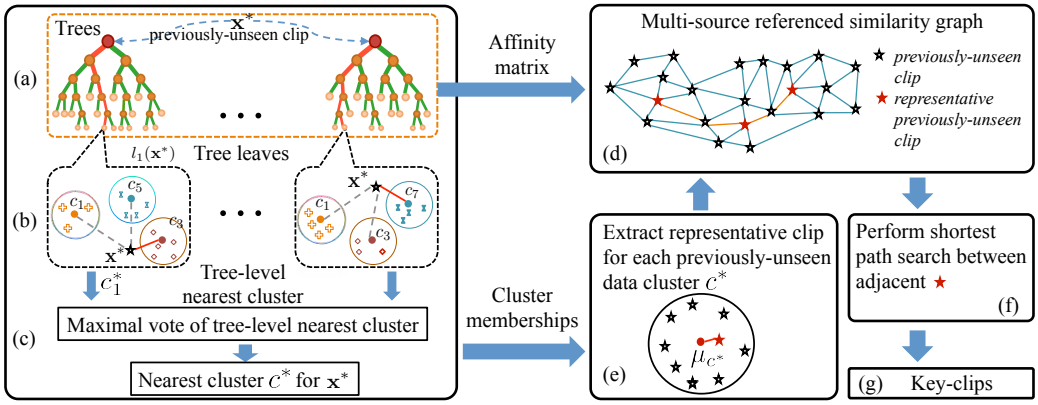


Figure 5.3: The pipeline of our multi-source referenced key-clips detection algorithm. (a) Channel a clip $\mathbf{x}^*$ into MSC-trees. (b) Search tree-level nearest clusters of $\mathbf{x}^*$, hollow circle denotes cluster. (c) Predict the final nearest cluster. A red $\star$ depicts a representative previously-unseen clip.

5.2 Semantic Video Summarisation

In Section 5.1 we presented multi-source data clustering by learning a Multi-Source Clustering Forest (MSC-Forest), resulting in a consistent cluster formation. Once this multi-source model is learned, it can be deployed for semantic video summarisation. Specifically, we follow the established approach of summarising videos by clustering (Truong and Venkatesh, 2007) but with the introduction of two noticeable differences in our method.

Firstly, our *video summary is multi-source referenced*. Specifically, the MSC-Forest is trained on heterogeneous sources, its optimised split functions $\{\mathrm{h}\}$ (Equation (2.1)) therefore implicitly capture the complex multi-source structures. When one deploys the trained model for content summarisation of previously-unseen video data, the model only needs to take visual inputs without any non-visual data sources. And yet it is able to induce video content partitions that not only correspond to visual feature similarities, but also are consistent with meaningful non-visual semantic interpretations. Secondly, our *video summary is automatically tagged* as the result of model inference. This is made possible through exploiting the non-visual data distributions associated with the discovered clusters on the training data (see Equation (5.2) and Figure 5.2(d)). Below we discuss the details of generating a semantic video summary.

5.2.1 Key-Clip Extraction and Composition

Suppose we are given a previously-unseen surveillance video footage without meta-data tagging/script. The video is pre-processed by segmenting it into a set of n^* either overlapping or non-overlapping short clips $\{\mathbf{x}_i^*\}_{i=1}^{n^*}$ with equal duration. Our aim is to first assign cluster mem-

bership to each previously-unseen clip using the trained multi-source model, and then select key-clips from the resulting clusters[2]. The chosen key-clips are then chronologically ordered to construct a video summary.

Clustering previously-unseen video clips Inferring cluster memberships of previously-unseen clips is an intricate task. A straightforward method is to assign cluster membership by identifying the nearest cluster $c^* \in C$ to a sample $\mathbf{x}^*$, where C represents all clusters we discovered in Section 5.1.2. However, we found this hard cluster assignment strategy susceptible to outliers in C and source noise. To mitigate this problem, we consider an alternative approach by utilising the MSC-Forest tree structures for soft cluster assignment. This is more robust to either source noise or outliers.

Figure 5.3 depicts the soft cluster assignment pipeline. First, we trace the leaf $l_t(\mathbf{x}^*)$ of each tree t where $\mathbf{x}^*$ falls by channelling $\mathbf{x}^*$ into the tree (Figure 5.3(a)). This step is critical as it establishes a connection for $\mathbf{x}^*$ with an appropriate training subset $X_{l_t(\mathbf{x}^*)}$ using the split functions $\{\mathrm{h}\}_t$ optimised by multi-source data. Here, $X_{l_t(\mathbf{x}^*)}$ represents the set of training samples associated with $l_t(\mathbf{x}^*)$. The set is consistent with $\mathbf{x}^*$ both visually and semantically since they encompass identical response w.r.t. $\{\mathrm{h}\}_t$.

Second, we retrieve the cluster membership $C_t = \{c_i\} \subset C$ of $X_{l_t(\mathbf{x}^*)}$, against which we search for the tree-level nearest cluster c_t^* for $\mathbf{x}^*$ (Figure 5.3(b)) via

$$c_t^* = \mathrm{argmin}_{c \in C_t} ||\mathbf{x}^* - \mu_c||, \tag{5.6}$$

with t the tree index, and μ_c the centroid of cluster c, estimated as

$$\mu_c = \frac{1}{|X_c|} \sum_{\mathbf{x}_i \in X_c} \mathbf{x}_i, \tag{5.7}$$

where X_c represents the set of training samples in c. Performing nearest cluster search within C_t rather than the whole cluster space C brings a key benefit: since the search space is constrained by MSC-tree, it is more meaningful and also less noisy than the entire space C, leading to more accurate c_t^* estimation.

Once we obtain all tree-level nearest clusters from all the trees in the forest, $\{c_t^*\}_{t=1}^{\tau_{\text{clust}}}$, the

[2] It is worth noticing that the purpose of this clustering step is completely different from the multi-source data clustering during model training, as presented in Section 5.1.2. The latter is a component of our multi-source model training pipeline (Figure 5.2), whilst the former aims at revealing the latent structure over testing data for video summarisation.

final nearest cluster c^* is obtained as the one with maximal votes from all the trees (Figure 5.3(c))

$$c^* = \max \{c_t^*\}_{t=1}^{\tau_{\text{clust}}} \tag{5.8}$$

By repeating the above steps on all previously-unseen clips $\{\mathbf{x}_i^*\}_{i=1}^{n^*}$, we obtain their cluster labels as $C^* = \{c_i^*\}_{i=1}^{n^*}$ (Figure 5.3(e)).

Extracting key-clips With the assigned cluster memberships C^* on all previously-unseen clips, the key-clip of a previously-unseen video data cluster c^* can be represented by the representative previously-unseen clip $\mathbf{x}_r^*$ that is closest to the cluster centroid μ_{c^*} (Figure 5.3(e)). Concatenating these key-clips chronologically establishes a visual summary. Such a summary, however, is likely to be discontinuous in preserving visual context therefore non-smooth visually due to abrupt changes between adjacent key-clips. To enforce some degrees of smoothness in the visualisation of video summary whilst minimising redundancy, we adopt a shortest path strategy (Boccaletti et al., 2006) to induce an optimal path between two temporally-adjacent representative $\mathbf{x}_r^*$ on a graph $\ddot{G}$. This approach produces a visually more coherent video summary whilst discards as much redundancy as possible.

More precisely, we construct a graph $\ddot{G} = (\ddot{V}, \ddot{E})$, where $\ddot{V}$ and $\ddot{E}$ indicate the set of previously-unseen video clip vertices and edges (Figure 5.3(d)). The weights of edges can be efficiently estimated using Equation (3.4) and (3.5). Note that the graph $\ddot{G}$ is also multi-source referenced since it is derived from our multi-source MSC-Forest model. We then perform shortest path search between temporally-adjacent $\mathbf{x}_r^*$ on $\ddot{G}$ (Figure 5.3(f)) and all the samples that lie on the shortest paths compose the final key-clip set K (Figure 5.3(g)).

5.2.2 Video Tagging

Summarising video with high-level interpretation requires plausible meaningful content inference from video data $\mathbf{x}^*$. We derive a tree-structure aware tag inference algorithm capable of predicting tag types same as training non-visual data, based on the learned MSC-Forest and discovered training data clusters. Specifically, we first obtain the tree-level nearest cluster c_t^* of a previously-unseen sample $\mathbf{x}^*$ using Equation (5.6). Second, the $p(y_i|c_t^*)$ associated with c_t^* is utilised as the tree-level non-visual tag estimation for the i-th non-visual data type. To achieve a smooth prediction, we average all $p(y_i|c = c_t^*)$ obtained from individual trees as

$$p(y_i|\mathbf{x}^*) = \frac{1}{\tau_{\text{clust}}} \sum\nolimits_{t=1}^{\tau_{\text{clust}}} p(y_i|c_t^*). \tag{5.9}$$

Algorithm 3: Infer non-visual tags of previously-unseen clips.

Input: A previously-unseen clip $\mathbf{x}^*$, a trained MSC-Forest, training data clusters C;

Output: Predicted tag $\hat{y}_i$;

Initialisation:

  Compute $p(y_i|c)$ for each training data cluster (Equation (5.2));

  Compute cluster centroid μ_c (Equation (5.7));

Non-Visual Tag Inference:

for $t \leftarrow 1$ **to** τ_{clust} **do**

    Trace the leaf $l_t(\mathbf{x}^*)$ where $\mathbf{x}^*$ falls (Figure 5.3(a));

    Retrieve the training samples $X_{l_t(\mathbf{x}^*)}$ associated with $l_t(\mathbf{x}^*)$;

    Obtain the clusters $C_t = \{c_i\} \subset C$ of $X_{l_t(\mathbf{x}^*)}$;

    Search the tree-level nearest cluster c_t^* of $\mathbf{x}^*$ within C_t (Equation (5.6));

end

Estimate tag distribution $p(y_i|\mathbf{x}^*)$ (Equation (5.9));

Compute the final tag $\hat{y}_i$ (Equation (5.10)).

The final tag $\hat{y}_i$ for the ith non-visual type is obtained as

$$\hat{y}_i = \text{argmax}_{y_i}\, p(y_i|\mathbf{x}^*). \tag{5.10}$$

With the above steps, we can estimate all m non-visual tags $\hat{y}_i$s with $i \in \{1,\ldots,m\}$. The procedure of our tagging algorithm is summarised in Algorithm 3.

Given the extracted key-clips K and automatic assignment of non-visual tags (Equation (5.10)), we can now construct a video summary by chronologically concatenating each clip $\mathbf{x}^* \in K$ with smooth inter-clip transition, e.g. cross-fading, and labelling each clip with their inferred tags.

5.3 Datasets and Experimental Settings

Datasets We conducted experiments on two datasets collected from publicly accessible web-cams that feature an outdoor and an indoor scene respectively: (1) the TImes Square Intersection (TISI) dataset, and (2) the Educational Resource Centre (ERCe) dataset[3]. There are a total of 7324 video clips spanning over 14 days in the TISI dataset, whilst a total of 13817 clips were collected across a period of two months in the ERCe dataset. Each clip has a duration of 20

[3] Datasets available: www.eecs.qmul.ac.uk/%7Exz303/download.html

Table 5.1: Details of datasets. FPS = frames per second.

Dataset	Resolution	FPS	# Training Clip	# Deployment Clip
TISI	550×960	10	5819	1505
ERCe	480×640	5	9387	4430

seconds. The details of the datasets and training/deployment partitions are given in Table 5.1. Example frames are shown in Figure 5.4.

The TISI dataset is challenging due to severe inter-object occlusion, complex behaviour patterns, and large illumination variations caused by both natural and artificial lighting sources at different day time. The ERCe dataset is non-trivial due to a wide range of physical events involved that are characterised by large changes in environmental set-up, participants, crowdedness, and intricate activity patterns.

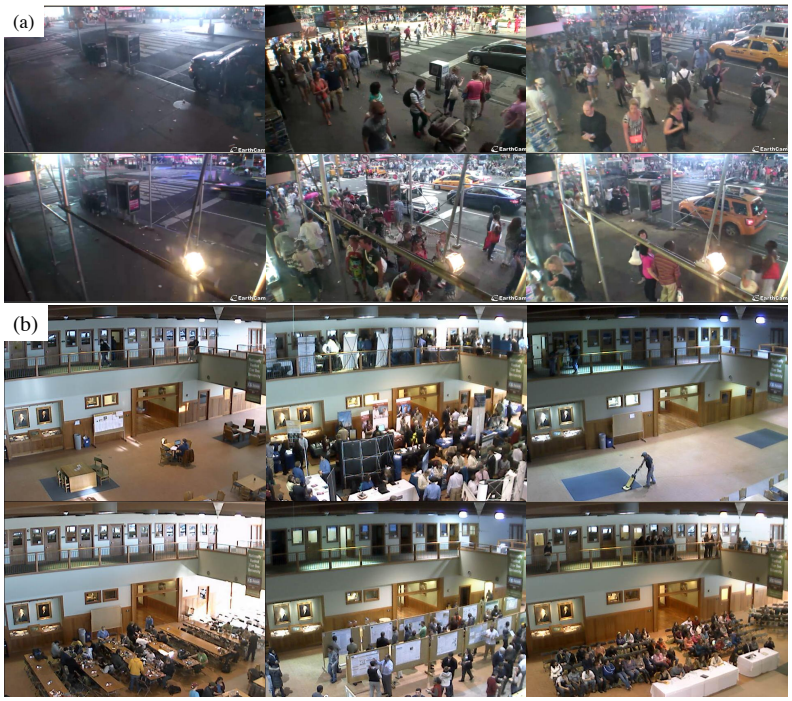


Figure 5.4: Examples of the (a) TISI and (b) ERCe datasets.

Visual and non-visual sources We extracted the following set of visual features for representing visual content in each clip: (a) colour features including RGB and HSV; (b) local texture features based on Local Binary Pattern (LBP) (Ojala et al., 2002); (c) optical flow; (d) holistic features of the scene based on GIST (Oliva and Torralba, 2001); and (e) person and vehicle[4]

[4]No vehicle detection on the ERCe dataset.

detection (Felzenszwalb et al., 2010).

We collected 10 types of non-visual sources for the TISI dataset: (a) weather data extracted from the WorldWeatherOnline with 9 elements: temperature, weather type, wind speed, wind direction, precipitation, humidity, visibility, pressure, and cloud cover; (b) traffic speed data from the Google Maps with 4 levels of traffic speed: very slow, slow, moderate, and fast. For the ERCe dataset, we collected data from multiple independent on-line sources about the time table of campus events including: No Scheduled Event (NoEvt), Cleaning (Cln), Career Fair (CrF), Gun Forum Control and Gun Violence (GunFrm), Group Studying (GrStd), Scholarship Competition (SchlCpt), Accommodative Service (AcmSvc), Student Orientation (StdOrt).

Note that other visual features and non-visual data types can be considered without altering the training and inference methods of our model in that the MSC-Forest model is capable of coping with different families of visual features as well as distinct types of non-visual sources.

Baselines To evaluate the proposed method for multi-source video clustering and tag inference, we compared the Visual + Non-Visual + MSC-Forest (*VNV-MSC-Forest*) model against the following baseline models:

1. *VO-Forest*: a conventional forest (Breiman, 2001) trained with visual feature vectors alone, to demonstrate the benefits from using non-visual sources[5]. Note that although no non-visual data is utilised for training the random forest, we still assume the availability of non-visual data after the video clusters are formed for the purposes of measuring the clustering coherence and video tagging.
2. *VNV-Kmeans*: k-means (Jain, 2010) using concatenated vectors of visual and non-visual features, to highlight the heteroscedasticity and dimensionality discrepancy problem caused by heterogeneous visual and non-visual data.
3. *VNV-Forest*: a conventional forest (Breiman, 2001) trained with concatenated visual and non-visual feature vectors, to compare the effectiveness of MSC-Forest that exploits non-visual data during forest formation.
4. *VNV-AASC*: a state-of-the-art multi-source spectral clustering method (Huang et al., 2012) learned by treating each type of visual or non-visual feature as an individual source, to demonstrate the superiority of MSC-Forest in handling diverse data representations and

[5] Evaluating a forest that takes only non-visual inputs is not possible, since non-visual data is not available for previously-unseen video footages.

correlating multiple sources.

5. *VNV-COP-Mahal*: a state-of-the-art Mahalanobis distance metric learning method (Xing et al., 2002) using both data features and two types of pairwise constraints, i.e. must-links: the two linked samples are in the same cluster; and cannot-links: the two linked samples are from two different clusters. In our multi-source data context, these pairwise constraints are generated from all non-visual data sources. Specifically, first, we computed individual similarity matrices from each non-visual source and averaged them for getting the fused pairwise similarity measure between video samples. The top-k highest and lowest pairwise similarity values were then used to generate must-links and cannot-links respectively. We set $k = \eta \times n \times (n-1) \times 10^{-5}$ where n is the number of training samples, whilst η was cross-validated in a range between 1 and 10, (i.e. k lies in $[339, 3390]$ on TISI, $[881, 8810]$ on ERCe), and the best results were utilised for comparison in our evaluation. Once the Mahalanobis distance metric was learned from both the visual feature data and the generated pairwise links using the algorithm proposed in (Xing et al., 2002), COP-Kmeans (Wagstaff et al., 2001) was employed along with pairwise links as well as the learned metric to obtain the final clusters of video data.
6. *VNV-DAKM* (Jones and Shao, 2014): a state-of-the-art dual assignment clustering method, Dual Assignment k-Means (DAKM), which is capable of performing two co-occurring clustering tasks simultaneously, while exploiting the correlation information to enhance both clusterings. In this multi-source context, visual and non-visual features are considered as two different views of clips. By exploiting their inherent correlation, DAKM attempts to improve the clustering on visual data while considering the semantic information encoded in the non-visual data. The parameter λ (in Equation (6) for controlling the uniformity of matrix $R^{'}$) was cross-validated in the range of $\{1, 10, 100, 1000\}$, and the best results were selected and compared.
7. *VNV-MSC-Forest-hard*: a variant of our model using hard cluster assignment strategy for inferring tags of previously-unseen samples (Section 5.2.2), to highlight the effectiveness of the proposed tree structure based tag inference algorithm.
8. *VT-MSC-Forest*: a variant of our model using only temporal information and visual data. In order to show the exact effectiveness of exploiting non-visual data, the weight ratio between visual data and time retains the same as in VNV-MSC-Forest with the only dif-

ference of discarding non-visual data during model training.

9. *VPNVρ-MSC-Forest*: a variant of our model but with $\rho\%$ of training samples having arbitrary number of missing non-visual types, to evaluate the robustness of MSC-Forest in coping with partial/missing non-visual data.

Implementation details The clustering forest size τ_{clust} was set to 1000, including both the conventional forest and the proposed MSC-Forest. We observed a slight increase in performance given a larger forest size, which agrees with (Criminisi and Shotton, 2012). The training set X_t of the t-th MSC-tree was obtained by performing random selection with replacement from the augmented data space (Figure 2.3(b)). We set $d_{\text{try}} = \sqrt{d}$ with d the data feature dimension (Equation (2.2)). This is typically practised (Breiman, 2001). We employed axis-aligned data separation (Criminisi and Shotton, 2012) as the test function for node splitting. This cluster number was discovered automatically using the method presented in (Zelnik-manor and Perona, 2004). We set the same number of clusters across all compared methods. For each dataset, $\sim 75\%$ out of the total data was utilised for model training, and the remaining was reserved for testing. Additional previously-unseen video data was collected from the Time Square Intersection scene on a separate day for video summarisation.

5.4 Experiments and Evaluations

5.4.1 Evaluation on Multi-Source Data Structure Discovery

To evaluate the effectiveness of different clustering models for multi-source video clustering, we compared the quality of their clusters formed on the training dataset. For determining clustering quality, we quantitatively measured the mean entropy (Zhao and Karypis, 2004) of non-visual distributions $p(y_i|c)$ (Equation (5.2)) associated with training data clusters to evaluate how coherent video content are partitioned, assuming all methods have access to non-visual data during the entropy computation.

It is evident from Table 5.2 that our VNV-MSC-Forest achieves the best cluster purity on both datasets[6]. Despite that there are gradual degradations in clustering quality when we increase the non-visual data missing proportion, overall the VNV-MSC-Forest model copes well with partial/missing non-visual data. With no aid of non-visual tag information, VT-MSC-Forest forms much worse clusters. Whilst the superiority of VT-MSC-Forest over VO-Forest suggests the

[6] VNV-MSC-Forest-hard shares the same clusters as VNV-MSC-Forest.

Figure 5.5: Qualitative comparison on cluster quality on TISI. A key frame of each video is shown. Numbers in brackets: First: the number of clips with sunny weather; Second: the total number of clips in a cluster. The frames inside the red boxes are inconsistent clips in a cluster.

effectiveness of temporal information with MSC-Forest. Inferior performance of VO-Forest to VNV-MSC-Forest suggests the importance of learning from auxiliary non-visual sources. Nevertheless, not all methods perform equally well when learning from the same visual and non-visual sources: the Kmeans, AASC, and COP-Mahal perform much poorer in comparison to MSC-Forest. The results suggest the proposed joint information gain criterion (Equation (5.1)) is more effective in handling heterogeneous data than the conventional clustering models.

For qualitative comparison, we show examples in Figure 5.5 using the TISI dataset for detecting 'sunny' weather. It is evident that only VNV-MSC-Forest is able to provide coherent video grouping, with only slight decrease in clustering purity given partial/missing non-visual data. Other methods including VNV-AASC result in a large cluster either leaving out some relevant clips or including many non-relevant ones, with most of them under the influence of strong artificial lighting sources. These non-relevant clips are visually 'close' to sunny weather, but semantically not. The VNV-MSC-Forest model avoids this mistake by correlating both visual and non-visual sources in an information theoretic sense.

Table 5.2: Compare cluster purity in mean entropy. Lower is better.

Dataset	TISI		ERCe
$p(\mathbf{y}\|c)$	traffic speed	weather	event
VO-Forest (Breiman, 2001)	0.8675	1.0676	0.0616
VNV-Kmeans (Jain, 2010)	0.9197	1.4994	1.2519
VNV-Forest (Breiman, 2001)	0.8611	1.0889	0.0811
VNV-AASC (Huang et al., 2012)	0.7217	0.7039	0.0691
VNV-COP-Mahal (Xing et al., 2002)	0.8523	1.2301	1.0685
VNV-DAKM (Jones and Shao, 2014)	0.7454	0.8088	0.6531
VT-MSC-Forest	0.7275	0.9577	0.0580
VNV-MSC-Forest	0.7262	**0.6071**	**0.0024**
VPNV10-MSC-Forest	**0.7190**	0.6261	**0.0024**
VPNV20-MSC-Forest	0.7283	0.6497	0.0090

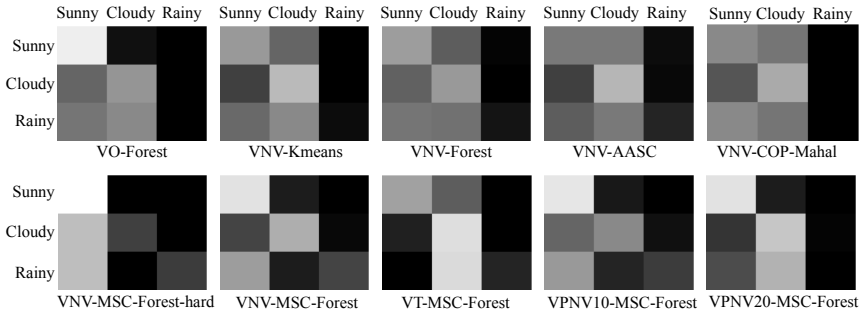


Figure 5.6: Weather tagging confusion matrices (TISI dataset).

5.4.2 Evaluation on Video Tagging

Generating video summary with semantic interpretations requires accurate tag prediction. In this experiment we compared the performance of different methods in inferring semantic tags given previously-unseen clips extracted from long videos. The proposed tagging algorithm (Section 5.2.2) is used for VO-Forest, VT-MSC-Forest, VNV-MSC-Forest, and VPNV10/20-MSC-Forest, whilst nearest neighbour (NN) strategy for the others. For quantitative evaluation, we manually annotated 3 weather conditions (sunny, cloudy and rainy) and 4 traffic speeds on TISI previously-unseen clips, whilst 8 event categories on ERCe previously-unseen clips.

Table 5.3: Comparison of tagging accuracy on the TISI dataset.

(%)	traffic speed	weather
VO-Forest (Breiman, 2001)	27.62	50.65
VNV-Kmeans (Jain, 2010)	37.80	43.14
VNV-Forest (Breiman, 2001)	34.95	43.81
VNV-AASC (Huang et al., 2012)	36.13	44.37
VNV-COP-Mahal (Xing et al., 2002)	26.22	40.03
VNV-DAKM (Jones and Shao, 2014)	**45.31**	42.87
VNV-MSC-Forest-hard	32.86	49.59
VT-MSC-Forest	35.99	54.47
VNV-MSC-Forest	35.77	**61.05**
VPNV10-MSC-Forest	37.99	55.99
VPNV20-MSC-Forest	38.05	54.97

Tagging video by weather and traffic conditions The experiment was conducted on the TISI outdoor dataset. It is observed that the performance of different methods (Table 5.3) is mostly in line with their performance in data clustering (Section 5.4.1). Poor result of tagging traffic conditions is yielded by VO-Forest. This suggests the significance of exploiting non-visual data during model training. It is also seen from Figure 5.6 that VNV-MSC-Forest not only outperforms other baselines in isolating the sunny weather, but also performs well in distinguishing visually ambiguous cloudy and rainy weathers. In contrast, both VNV-Kmeans and VNV-AASC mistake most of the 'rainy' scenes as either 'sunny' or 'cloudy', as they can be visually similar. DAKM produces the best accuracy on traffic speed while much poor on predicting weather. By examining the visual and non-visual, we found the plausible reason that weather data is more likely to be noisy, which may affect the correlation learning in DAKM and in turn leading to sub-optimal interaction between the two views. Interestingly, the poorest tagging results are obtained by VNV-COP-Mahal where non-visual data is alternatively used as side information for generating pairwise constraints over video samples. The potential reasons include (1) COP-Mahal assumes completely-accurate pairwise links, which however is largely invalid in our context due to the intrinsic noisy nature of non-visual data sources; (2) the errors in pairwise constraints can be propagated during the clustering process of COP-Kmeans and therefore is likely to further

Table 5.4: Comparison of tagging accuracy on the ERCe dataset.

(%)	NoEvt	Cln	CrF	GunFrm	GrStd	SchlCpt	AccSvc	StdOrt	*Average*
VO-Forest (Breiman, 2001)	79.48	39.50	94.41	74.82	92.97	82.74	00.00	60.94	65.61
VNV-Kmeans (Jain, 2010)	**87.91**	19.33	59.38	44.30	46.25	16.71	00.00	09.77	35.45
VNV-Forest (Breiman, 2001)	32.47	30.25	65.46	45.77	41.25	33.15	13.70	33.59	36.96
VNV-AASC (Huang et al., 2012)	48.51	45.80	79.77	84.93	96.88	89.40	21.15	38.87	63.16
VNV-COP-Mahal (Xing et al., 2002)	41.98	**71.43**	54.61	15.07	21.88	00.00	00.24	00.00	25.65
VNV-DAKM (Jones and Shao, 2014)	65.35	25.63	62.34	44.49	53.75	18.61	7.21	14.06	36.43
VNV-MSC-Forest-hard	81.25	41.60	70.07	60.48	84.22	82.88	10.82	47.85	59.89
VT-MSC-Forest	57.43	70.17	91.45	79.96	**99.22**	90.08	00.00	43.75	66.50
VNV-MSC-Forest	55.98	41.28	**100.0**	83.82	97.66	99.46	**37.26**	88.09	75.69
VPNV10-MSC-Forest	47.96	46.64	**100.0**	**85.29**	97.66	**99.73**	**37.26**	**92.38**	75.87
VPNV20-MSC-Forest	55.57	46.22	**100.0**	**85.29**	95.78	99.59	37.02	88.09	**75.95**

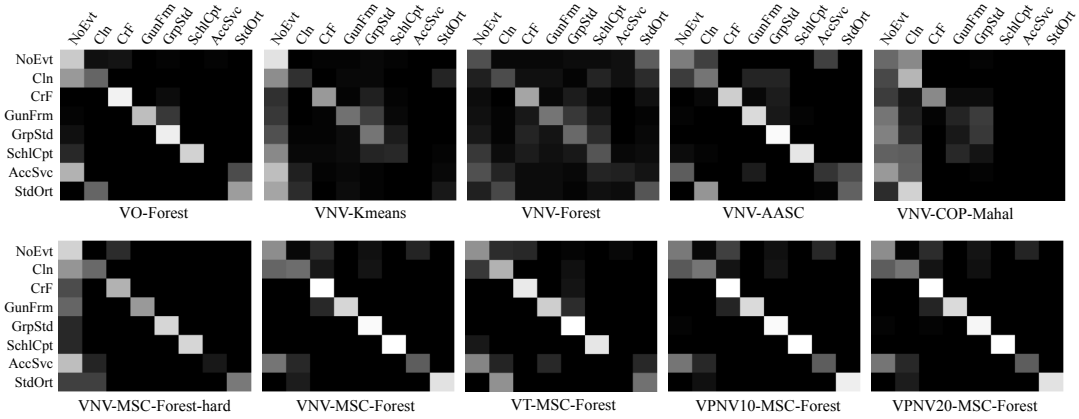


Figure 5.7: Event tagging confusion matrices (ERCe dataset).

worsen the cluster solution and finally the tagging accuracy. This reflects the significant difficulty of jointly learning inherently heterogeneous and inaccurate visual and non-visual data as aforementioned, and in turn the advantages of the proposed joint information gain formulation over existing competitive algorithms.

Tagging video by activity events Tagging semantic events was tested using the ERCe dataset. By VO-Forest, poor results (Table 5.4 and Figure 5.7) are obtained especially on 'Accommodation Service', which involves only subtle activity patterns, i.e. students visiting particular rooms, suggesting using visual data alone is not sufficient to detect such visually subtle events. VT-MSC-Forest over-fits to 'Cleaning' event, therefore performs poorly on 'Student Orientation' event.

Due to the typical high-dimension of visual sources compared to non-visual data, the latter is often overwhelmed by the former in representation. VNV-Kmeans severely suffers from this problem as its most predictions are biased to 'No Scheduled Event' that is more common

and frequent visually. This suggests that this distance-based clustering is poor in handling the heteroscedasticity and dimension discrepancy problems in learning heterogeneous data. VNV-AASC attempts to circumvent these problems by seeking for an optimal combination of affinity matrices derived independently from distinct data sources. However this is proved challenging, particularly when each source is inherently noisy and inaccurate. Similar to the observation on TISI, DAKM obtains poor results in distinguishing complex social events mainly because the corresponding event non-visual data is not well synchronised with visual data, which results in inaccurate mutual correlation and finally negative interaction to each other. As an alternative way of utilising non-visual data, VNV-COP-Mahal yields again the lowest overall accuracy. This further shows the unsuitability of COP-Mahal in learning ambiguous heterogeneous data due to its stringent assumption on the availability of accurate and reliable pairwise links and the lack of noisy data handling mechanism. In contrast, the proposed MSC-Forest correlates different sources via a joint information gain criterion to effectively alleviate these problems, leading to more robust and accurate tagging performance. Again, VPNV10/20-MSC-Forest perform comparably to VNV-MSC-Forest, further validating the robustness of MSC-Forest in tackling partial/missing non-visual data with the proposed adaptive weighting mechanism (Section 5.1.1).

Interestingly, in some cases, VPNV10/20-MSC-Forest models even outperform VNV-MSC-Forest slightly. We observe that this can be caused by missing noisy non-visual data, which may lead to better results. Overall, the performance difference is marginal and the results demonstrate that MSC-Forest provides stable tagging results across both datasets.

Evaluating α sensitivity We analyse the relative significance of visual data against non-visual and temporal data by varying its weight α_v (Equation (5.1)) in MSC-Forest during model training. The average tagging accuracy is utilised as performance measure criterion. It is observed from Figure 5.8 that setting $\alpha_v = 0.5$ achieves satisfactory results for both datasets. This observation suggests that visual and non-visual data are almost equally informative. This setting of α is adopted throughout our experiments.

5.4.3 Evaluation on Semantic Video Summarisation

In this experiment, we follow the method described in Section 5.2, and show that the learned model MSC-Forest can be easily extended to produce compact yet meaningful video summary of previously-unseen video footage, e.g. from the Time Square Intersection scene, with automat-

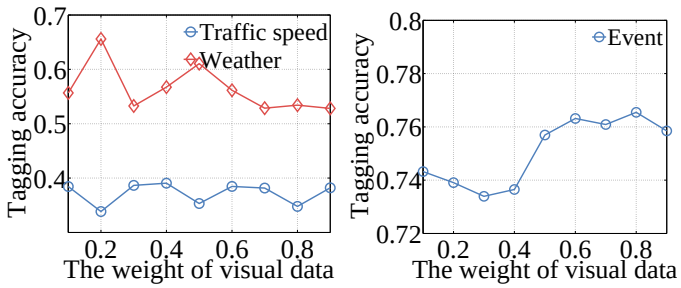


Figure 5.8: The average tagging accuracy against varying visual data weight α_v in Equation (5.1).

ically generated tags. Despite captured from the same scene as the TISI dataset, this previously-unseen video is challenging in that it contains a number of events not seen before (e.g. scaffolding event), with very different weather and traffic conditions. It is interesting to examine how well the multi-source model could generalise for drawing meaningful summarisation given such unexpected disparities.

A Quantitative Evaluation on Summary Quality

Measuring the quality of video summary quantitatively is non-trivial since there is no formal definition in the literature. In this study, we employ a *coverage* metric – an ideal summary should cover as many events of interest as possible[7]. More precisely, given a video summary V^{smr}, its coverage is defined as $\widetilde{c} = \frac{n_{\text{cvd}}}{n_{\text{all}}}\left(\frac{\max_i |V_i^{\text{smr}}|}{|V^{\text{smr}}|}\right)$, where n_{cvd} and n_{all} represent the number of covered and all events of interest, respectively. The $|V^{\text{smr}}|$ is the length of the current summary, whilst $\max_i |V_i^{\text{smr}}|$ represents the maximum length of all comparative synopses. The term $\left(\frac{\max_i |V_i^{\text{smr}}|}{|V^{\text{smr}}|}\right)$ thus penalises a summary with longer length. Higher coverage is better, implying lower redundancy.

In order to generate unbiased ground truth of event of interest, we asked 10 annotators to watch the previously-unseen video carefully and label each video clip with arbitrary event tags. Although these event tags were produced independently in a somewhat subjective manner, the repetition of similar tagging among different annotators is high, e.g. most annotators labelled 'unloading scaffolding tubes', 'policemen on-duty', as events of their interest. Thus, we formed the ground truth with events that were agreed by over 50% of the annotators. The final ground truth consists of 12 events (Figure 5.9).

Given the ground truth, we compared the quality of summary generated using the proposed multi-source MSC-Forest with the baselines: (1) Uniform-Sampling: a straightforward way of

[7]The event of interest is analogous to important objects/regions in (Lee et al., 2012).

summarising video by uniformly sampling video clips over time, assuming key events are distributed evenly (Truong and Venkatesh, 2007; Lee et al., 2012). (2) Sufficient-Change: a type of classical summarisation strategy generic to video category (Zhang et al., 1997; Kim and Hwang, 2002; Truong and Venkatesh, 2007). The idea is to select the clip significantly different from the previous key clip e.g. using threshold based strategy and thus the extracted key clips may be of great diversity and complete. The threshold can be estimated based on the number of key clips. For the distance metric, we adopt L1-norm and L2-norm to measure pairwise similarity between clips in our experiment. (3) VO-Forest: the conventional Forest (Breiman, 2001) that exploits visual features alone. For VO-Forest and MSC-Forest, we applied the summarisation pipeline described in Section 5.2 for summary composition. We generated the video summary by the remaining methods via setting a duration similar to the summary by MSC-Forest. Note that non-visual information are not available during the summarisation stage. Hence, for clustering based models, the quality of a summary essentially ties to the purity and coherency of video clusters discovered using different methods.

The results are shown in Figure 5.9 and Table 5.5. It is evident that the MSC-Forest model achieves higher event coverage than the baselines. This is in large due to the MSC-Forest's ability for latent data structure discovery (Section 5.4.1). To reveal concrete reasons on the summarising performance difference, for the same previously-unseen samples $\mathbf{x}^*$ with event of interest, e.g. parcel delivery, we compared the assigned clusters: c^*_{vnv} by our model and c^*_{vo} by VO-Forest. It is found that samples in c^*_{vnv} are visually consistent each other and the majority share some similarity with $\mathbf{x}^*$, e.g. someone standing at the edge of pathway; whilst cluster c^*_{vo} is much larger with no obvious visual commonality over its cluster members. Uniform-Sampling performs poorly since the assumption of uniform event distribution is often invalid. Significant-Change is inferior to our model since the visual data distance/similarity measure can be inaccurate and less meaningful due to the challenging semantic gap problem.

A User Study on Summary Quality

We conducted a user study to examine if the non-visual tags inferred using the MSC-Forest model could complement the unilateral perspective offered by pure visual summary alone. We showed two video summaries to 10 volunteers: (i) a pure visual summary, and (ii) the same summary but enriched with semantic tags inferred using the proposed multi-source model[8]. The

[8]The inferred non-visual tags include weather, traffic conditions, and typicality. The typicality tag, i.e. *usual* and *interesting*, of each clip, is computed based on the size of their assigned clusters (Figure 5.3(c)).

Table 5.5: Quantitative comparison of video summary. Length = clip number.

Method	Length	Event number	Coverage
Uniform-Sampling	28	3	25.9%
Sufficient-Change(L1)	29	2	16.7%
Sufficient-Change(L2)	29	4	33.3%
VO-Forest	21	3	34.5%
VNV-MSC-Forest(**Ours**)	28	**7**	**60.4**%

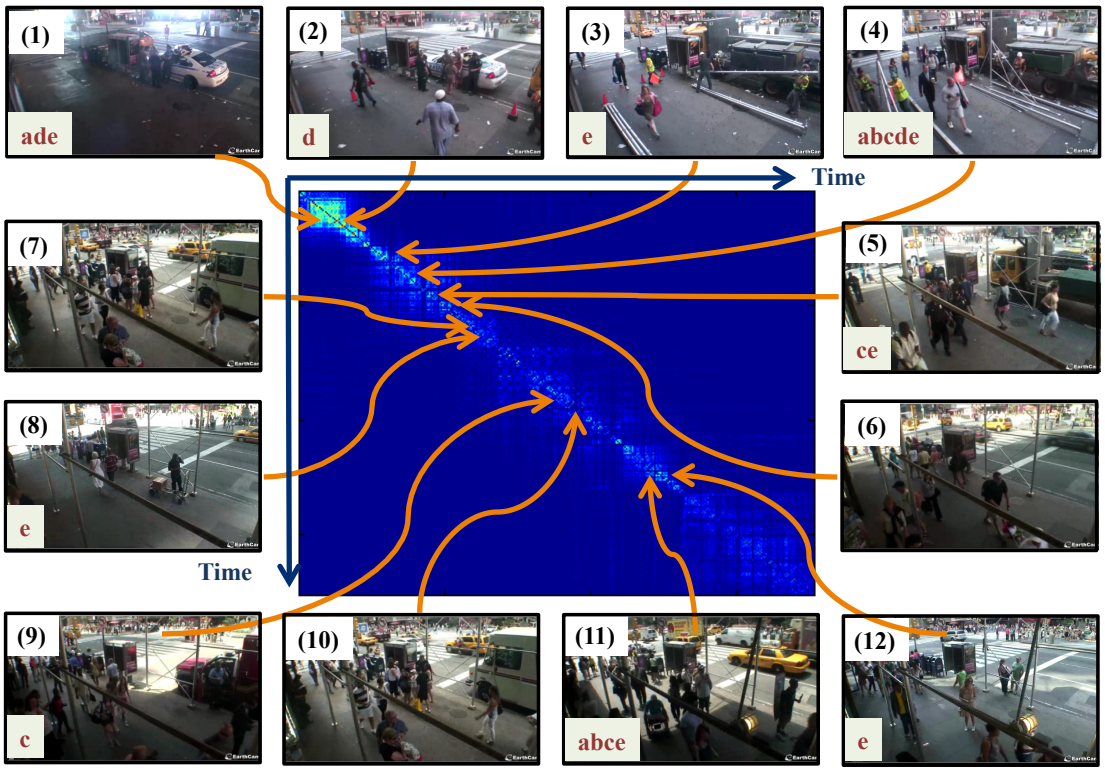


Figure 5.9: The multi-source affinity matrix constructed by our model, along with key frames corresponding to ground truth events of interest: (1) policemen on-duty, (2) blocking pathway, (3) workers unloading scaffolding tubes, (4)-(6) different stages of scaffolding, (7)(9)(10) van parking aside, (8) parcel delivery, (11)(12) loitering events. The event covered by some particular method is indicated on the left-bottom corner of key frame with their ID defined as: (a) Uniform-Sampling; (b) Sufficient-Change (L1); (c) Sufficient-Change (L2); (d) VO-Forest; (e) VNV-MSC-Forest.

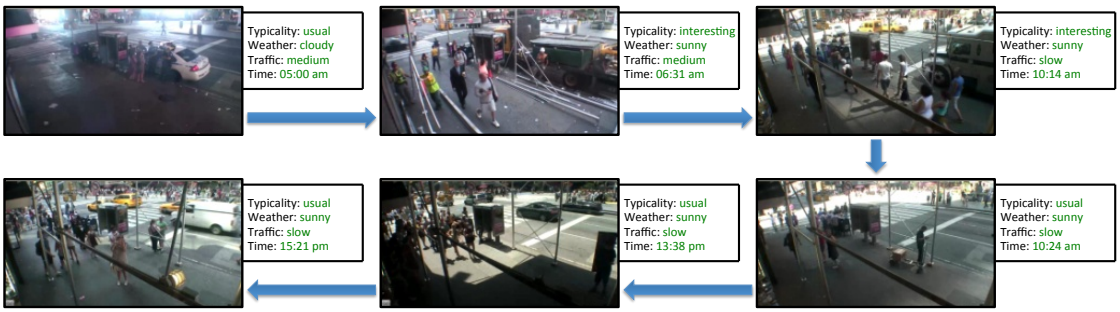


Figure 5.10: A storyboard version of our video summary enriched with non-visual tags.

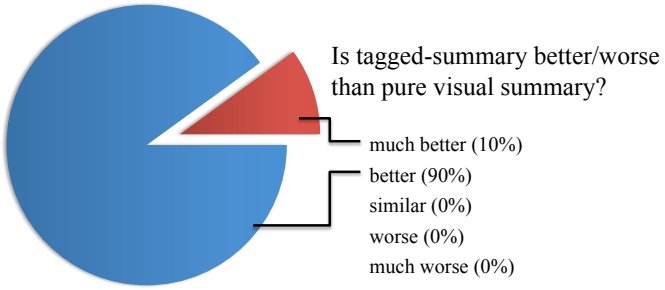


Figure 5.11: User study: tagged *versus* pure-visual summary.

tagged summary is shown in Figure 5.10. Each volunteer was asked to compare and rate the two summaries based on their preference. It is worth pointing out that passing the user test is challenging because providing additional non-visual tags to summary is not necessarily better than none. Tags that correlate poorly with visual context could even jeopardise user experience.

It is evident from Figure 5.11 that visual summary augmented with non-visual tags was well accepted by all participants over the conventional visual-only summary. A follow-up survey with the volunteers reveals several interesting reasons of their selection. Many volunteers found that the inferred non-visual tags were valuable in providing auxiliary context to achieve better global situational awareness. In particular, the tags helped them to 'connect the dots' and making sense of the previously-unseen (and likely unfamiliar) video footages. Some other volunteers credited the additional non-visual tags in focusing their attention on particular events, and helping them in spotting 'outliers' of interest.

This user study provides an independent means to analyse and validate the usefulness of visual summarisation with auto-tag inference of previously-unseen video footages without a priori semantics or meta-data, mostly typical of surveillance videos. It also shows the effectiveness of the proposed model for mapping multi-source non-visual information to unstructured and previously-unseen video data in automatic tagging and summarisation of the videos.

Clips assigned to the top 20% smallest clusters are treated as 'interesting'.

5.4.4 Multi-Source Model Visualisation

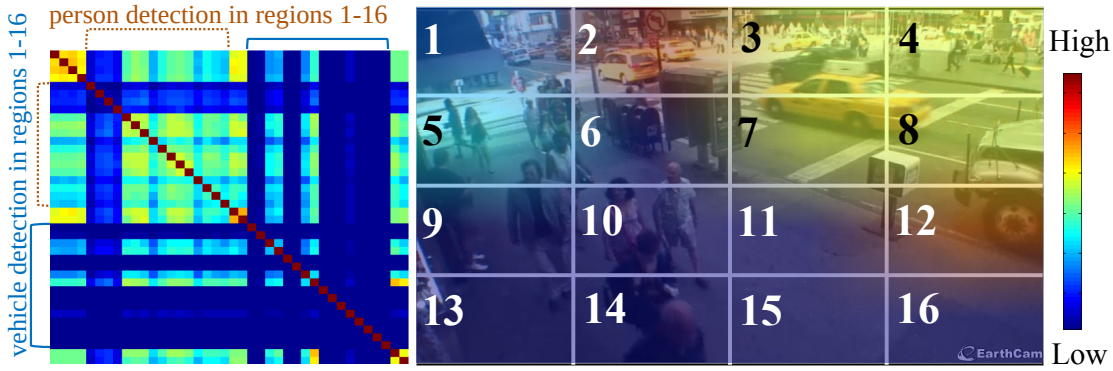
The superior performance of VNV-MSC-Forest can be better explained by examining more closely the capacity of MSC-Forest in uncovering and exploiting the intrinsic correlation among different visual sources and more critically among visual and non-visual sources. This indirect correlation among heterogeneous sources results in well-structured decision trees, subsequently leading to more consistent data clusters and more accurate semantics inference. The details of computing the multi-source correlation are presented in Section 5.1.3. Here we show an example multi-source correction revealed by our MSC-Forest for model visualisation purpose.

Intuitively, vehicle and person counts should correlate in a busy scene like TISI. Our MSC-Forest discovered this correlation (see Figure 5.12(a)), so the less reliable vehicle detection from distance against a cluttered background, could enjoy a latent support from more reliable person detection in regions 5-16 close to the camera view.

Moreover, visual sources also benefit from correlated support from non-visual data through our cross-sources information gain optimisation (Equation (5.1)). An example is the intuitive correlation between traffic speed and visual appearance, e.g. slow traffic speed often corresponds to crowded scenarios with a large quantity of pedestrians and vehicles whilst fast traffic speed to sparse people and cars. Such cross-source correlation can be captured by our MSC-Forest, as observed in Figure 10(b) that the vehicle detection responses over road area present a stronger interaction with traffic speed data than those on walk path where vehicles should not appear. In other words, vehicle detection features of road area are preferred over those on walk path in node splitting due to larger induced joint information gain (Equation (5.1)), which is clearly desired. This discovered correlation is further exploited by MSC-Forest during the node splitting optimisation process and thus facilitates the separation of different crowdedness levels of visual data. This leads to better clusters and eventually benefits video summarisation.

5.4.5 Computational Cost Analysis

We examined the computational costs for training the proposed MSC-Forest, in comparison to the conventional forests. Time is measured on a Windows PC machine with a dual-core CPU @ 2.66 GHz, 4.0GB RAM, with C++ implementation. Only one core is utilised for training each forest. We recorded the model training time under the same experimental setting as stated in Section 5.3. It is observed from Table 5.6 that the training cost of a MSC-Forest model is



(a) Correlations between visual data (b) Correlations between vehicle detection and traffic speed

Figure 5.12: The discovered multi-source correlations by our MSC-Forest on TISI.

Table 5.6: Random forest model training complexity. Lower is better. TT = Training Time (unit is second).

Dataset	TISI		ERCe	
-	TT	Φ^*	TT	Φ^*
VO-Forest	10306	109392	21831	359247
VNV-Forest	10646	108865	22015	359364
VNV-MSC-Forest	**8823**	**91316**	**7845**	**137620**

significantly lower than that of learning conventional forests. In particular, VNV-MSC-Forest records a reduced training time by 14.4% and 17.1% on TISI, and 64.1% and 64.4% on ERCe, when compared with VO-Forest and VNV-Forest, respectively. We observed similar trend on the model inference time.

The lower computational cost of MSC-Forest is owing to its shallow and balanced trees, thanks to the additional non-visual and temporal information during tree optimisation. To make this concrete, we showed in Table 5.6 the averaged tree fan-in $\Phi^* = \frac{1}{\tau_{\text{clust}}} \sum_{t=1}^{\tau_{\text{clust}}} \Phi(t)$ of different forest models. A forest with shallow and balanced trees tend to have a small Φ^* (see Section 5.1.1 for a discussion on tree fan-in). In addition, we also profiled the length of path (from root to leaf node) traversed by training samples. A shallow and balanced tree tends to have shorter path length. The distributions depicted in Figure 5.13 suggest that MSC-Forest has a shallower and more balanced tree topology than that of conventional forests. It is worth pointing out that despite the shallower structure, MSC-Forest outperforms other models in our clustering and tagging experiments.

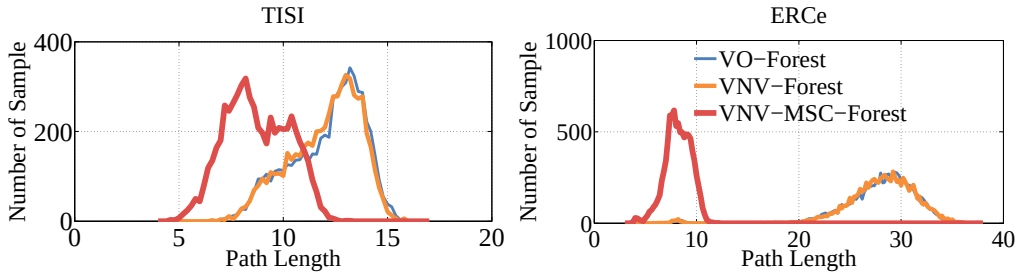


Figure 5.13: Comparing tree path length statistics. The same legend is used for both charts.

5.5 Summary

In this chapter, an unsupervised multi-source data structure discovery model is formulated and applied for semantic surveillance video summarisation. Specifically, we introduced a joint information gain function for discovering and exploiting latent correlations among independent heterogeneous visual and non-visual data sources. This formulation naturally copes with diverse types of data with different representations, distributions, and dimensions. Importantly, our model is capable of tolerating partial and missing non-visual data during model training, and allowing automatic tag inference on previously-unseen video footages and for video summarisation. Furthermore, the proposed joint optimisation encourages more compact decision trees, leading to more efficient model training and tag inference. Extensive comparative experiments have demonstrated the advantages of the proposed multi-source video clustering model over existing visual-only models, for both discovering latent video clusters and inferring non-visual tags on previously-unseen video footages. A comprehensive user study was carried out to validate independently the effectiveness of deploying the proposed model for generating contextually-rich and semantically-meaningful video summary. The proposed model is not limited to surveillance-type videos but can be generalised to other types of unstructured and un-tagged consumer videos or egocentric videos, if 3D camera motion-invariant features or egocentric features (Lee et al., 2012) are adopted.

The methods presented in the Chapters (3, 4, 5) are designed for data captured from a single camera view. Their learning schemes may not be appropriate to process video data captured from multiple cameras, much typical in video surveillance settings. Specifically, these algorithms are likely to suffer from the large cross-view variations in lighting condition, background clutter and occlusion. It is non-trivial for computing models to overcome these largely uncontrollable difficulties and challenges. The following chapter presents a model that is particularly formu-

lated for coping with multi-camera visual data through discriminative learning for quantifying the considerable difference and disparity in viewing conditions across camera views. The aim of multi-camera data structure discovery model is to identify person-specific structural patterns among a population across multiple distributed surveillance cameras.

Chapter 6

Person Identity Structure Discovery by Discriminative Selection in Video Ranking

---

The proceeding Chapters (3, 4, 5) describe a set of data structure discovery algorithms for visual data captured from a single camera view. The range from a single camera is limited in viewing field for video surveillance. This naturally necessities the deployment of multiple connected cameras in order to jointly monitor wide public areas, e.g. underground stations and airport terminals. In such multi-camera surveillance scenarios, an essential requirement is to associate people across disjoint camera views so as to uncover person-specific activity distribution/structures beyond the view range of single cameras. That is, it aims at finding person identity cluster structures among cross-camera visual data. This is also called person re-identification (ReID) in the literature.

Existing person ReID methods typically rely on single-frame imagery features (Gong et al., 2014a), whilst ignoring space-time information from image sequences often available in the practical surveillance scenarios. Single-frame (single-shot) based visual appearance matching is inherently limited for person re-identification in public spaces due to the challenging visual ambiguity and uncertainty arising from non-overlapping camera views where viewing condition changes can cause significant people appearance variations. On the other hand, psychology and physiology research studies suggest that human vision system's capability to recognise dynamic sequential targets, e.g. face identity, is superior than that to static observations such as still images. Bassili (1979) found that the movement of a sparse spatial arrange of white dots

over a black face surface, i.e. space-time facial feature representation allows superior expression recognition than stationary images, even though using a small number of features. Knight and Johnston (1997) showed that faces in photographic negatives can be more effectively recognised when shown moving than static. Moreover, Davis and Bobick (1997) demonstrated that actions encoded in blurred image sequences where little structure knowledge is presented in individual image frames can be easily recognised by human.

In this chapter, a person ReID approach is proposed for learning person discriminative space-time dynamic information from image sequences. Specifically, this model can automatically select the most discriminative video fragments from noisy and incomplete walking sequences of people from which more reliable space-time and appearance features can be computed, whilst simultaneously to learn a video ranking function for person re-identification.

The remainder of this chapter is structured as below. Section 6.1 provides the details of the proposed space-time sequence based ReID method. Datasets and experimental settings are described in Section 6.2. Followed are experiments and evaluations with comparison to contemporary gait recognition, holistic image sequence matching and state-of-the-art single-shot/multi-shot based re-identification methods in Section 6.3. Finally, a summary is presented in Section 6.4.

6.1 Discriminative Video Ranking

We formulate the person re-identification problem as a ranking problem (Prosser et al., 2010; Gong et al., 2014a). Although image sequences of people may provide intuitively richer content to learn discriminative information about individual's visual appearance when compared to a single still image widely used by existing person re-identification methods (Li and Wang, 2013b; Gray et al., 2007; Zheng et al., 2009; Loy et al., 2009), the availability of more (and often redundant) data poses additional challenges in model learning, e.g. more random inter-object occlusions and thus incomplete frames, arbitrary sequence duration and starting/ending postures. Moreover, human annotators may implicitly and unconsciously have the tendency to manually select carefully clearer and better-segmented person images for learning image-based re-identification models. On the other hand, automatically detected and tracked sequences of person bounding boxes in typical surveillance videos are inherently noisier and incomplete. Directly utilising *all* the sequence data for constructing re-identification models can easily result

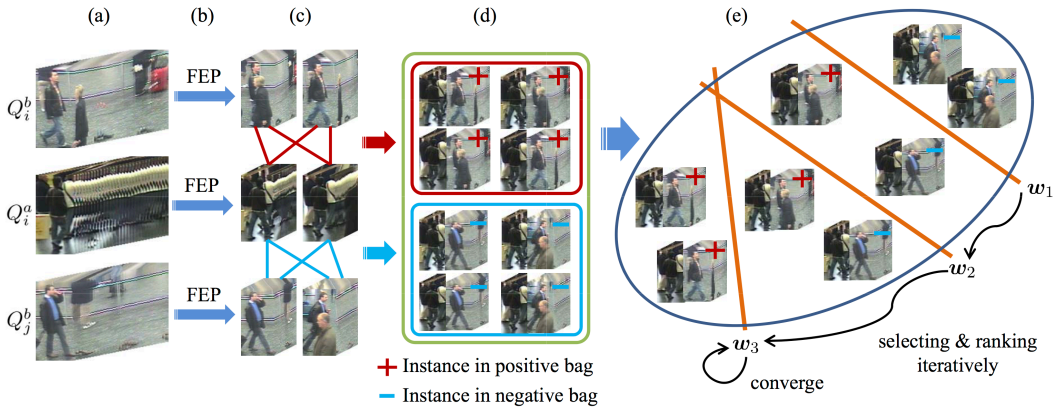


Figure 6.1: The training pipeline of the proposed discriminative video ranking framework. (a) Image sequences of training people, Q_i^a denotes the image sequence of person p_i from camera a (Section 6.1.1). (b)-(c) Generating candidate fragments for each sequence by motion energy profiling (Section 6.1.2). (d) Creating cross-view fragment pairs as positive and negative instances and pooling them into positive and negative bags respectively (Section 6.1.3). (e) Learning a sequence-based relative ranking model by simultaneously selecting and ranking iteratively discriminative fragment pairs (Section 6.1.3).

in unstable models, therefore is undesirable. A selection mechanism is required to be part of the learning model in order to optimally explore the redundant information available in sequence data.

In the context of relative ranking based re-identification model learning, it is non-trivial to automatically learn a robust discriminative ranking function from such contaminated image sequence data for person re-identification. Inherently, one needs to address the problem of how to mitigate the negative influence of unknown noisy observations, e.g. various types of occlusions and clutters in the background. This is more than solving the more common problem of misalignment over time in sequence matching. In this work, we formulate a discriminative re-identification model capable of simultaneously selecting and ranking informative video fragments from pairs of unregulated image sequences of people captured in two non-overlapping camera views. Our model not only mitigates unwanted data whilst exploring useful information for person re-identification from image sequences, but also requires no rigid sequence alignment as in the case of traditional methods, e.g. dynamic time warping. More specifically, our model is based on : (i) Video fragmentation by motion energy profiling (Figure 6.1(b)(c) and Section 6.1.2) ; (ii) Learning a sequence-based relative ranking function by simultaneously selecting and ranking cross-view video fragment pairs (Figure 6.1(d)(e) and Section 6.1.3) . Once learned, our model can then be deployed to re-identify previously unseen people given cross-view unregulated image sequences (Section 6.1.4). An overview diagram of the learning process of the proposed

approach is presented in Figure 6.1.

6.1.1 Problem Definition

Suppose we have a collection of image sequence pairs $\{(Q_i^a, Q_i^b)\}_{i=1}^n$, where Q_i^a and Q_i^b denote the image sequences of person p_i captured by two disjoint cameras a and b, and n the number of people in the training set. Each image sequence is defined as a set of consecutive frames I obtained by an independent person tracker, e.g. (Ben Shitrit et al., 2011; Hare et al., 2011): $Q = \{\mathbf{I}_1, \mathbf{I}_2, ...\}$, where $|Q|$ is not a constant as in typical surveillance videos, tracked person image sequences do not guarantee to have a uniform duration (arbitrary number of frames), nor the number of walking cycles and starting/ending postures.

For model training, we aim to learn a ranking function $f(Q^a, Q^b)$ of image sequence pairs that satisfies the the following ranking constraints:

$$f(Q_i^a, Q_i^b) > f(Q_i^a, Q_j^b), \forall i = \{1, ..., n\}, \forall j \neq i. \tag{6.1}$$

That is, the image sequence pair (Q_i^a, Q_i^b) of the same person p_i is constrained/optimised to have a higher ranking over any cross-view sequence pairing of person p_i and p_j where $j \neq i$.

Learning a ranking function *holistically without discrimination and selection* from pairs of unsegmented and temporally unaligned person image sequences will subject the learned model to significant noise and degrade any meaningful discriminative information contained in the image sequences. This is an inherent drawback of any holistic sequence matching approach, including those with dynamic time warping applied for non-linear mapping (see experiments in Section 6.3). Reliable human parsing/pose detection (Kanaujia et al., 2007) or occlusion detection (Xiao et al., 2006) may help, but such approaches are difficult to be scaled, especially with image sequences from crowded public scenes. The challenge is to learn a robust ranking model effective in coping with incomplete and partial image sequences by identifying and selecting discriminative/informative video fragments from each sequence suitable for extracting trustworthy fragment features. Let us first consider generating a pool of candidate video fragments for each image sequence, i.e. video fragmentation.

6.1.2 Video Fragmentation

Given the unregulated image sequences of people, it is too noisy to attempt holistically locating and extracting reliable discriminative fragment features from an entire image sequence. Instead,

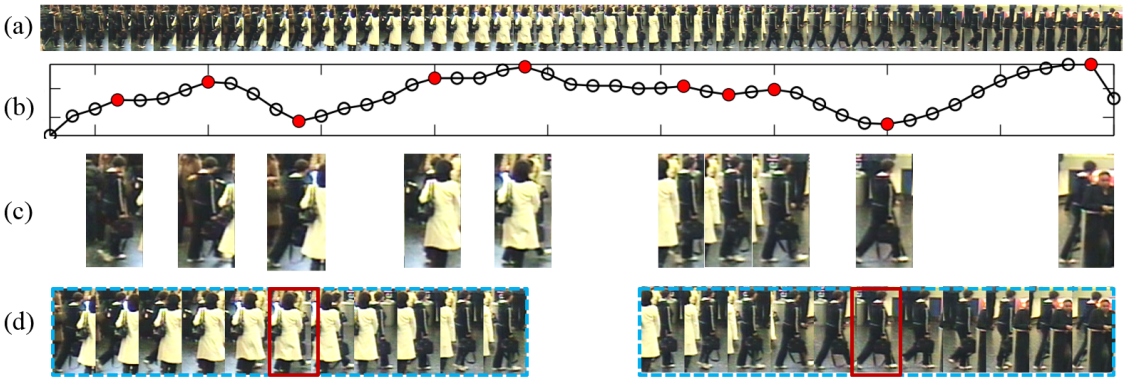


Figure 6.2: (a) A person sequence of 50 frames is shown, with the motion energy intensity of each component frame given in (b). The red dots in (b) denote the automatically detected local minima and maxima temporal landmarks in the motion intensity profile, of which the corresponding frames are provided at the vertically-aligned positions in (c). (d) Two example video fragments (shown every 2 frames) with the landmark frames highlighted by red bounding boxes.

we consider breaking down each image sequence into localised video fragments and generate a pool of video fragment candidates to allow for a learning model to automatically select the discriminative fragments (Section 6.1.3).

It can be observed that motion energy intensity induced by the activity of human muscles during walking exhibits regular periodicity (Waters and Morris, 1972). This motion energy intensity can be approximately estimated by optic flow computation. We call this Flow Energy Profile (FEP), see Figure 6.2. This FEP signal is particularly suitable to address our video fragmentation problem due to: (i) the local minima and maxima landmarks are likely to correspond to some characteristic gestures in a person's walking process, thus help in estimating these characteristic walking postures (e.g. one foot is about to land); (ii) relatively robust to changes in camera viewpoint. More specifically, given a sequence $Q = \{\mathbf{I}_1, \mathbf{I}_2, ...\}$, we first compute the optic flow field $(\mathbf{V}_x, \mathbf{V}_y)$ for each image frame $\mathbf{I}$. The flow energy e of $\mathbf{I}$ is defined as

$$e(\mathbf{I}) = \sum_{(i,j)\in U} \|[\mathbf{V}_x(i,j), \mathbf{V}_y(i,j)]\|_2, \tag{6.2}$$

where U is the pixel set of the lower body, e.g. the lower half of an image $\mathbf{I}$. The FEP Q_{fep} of an image sequence Q is then obtained as $Q_{\text{fep}} = \{e(\mathbf{I}_1), e(\mathbf{I}_2), ...\}$, which is further smoothed by a Gaussian filter to suppress noise.

Subsequently, we locate the local minima and maxima landmarks $\{t\}$ of Q_{fep} and for each landmark create a video fragment g by extracting the surrounding frames $g = \{\mathbf{I}_{t-l_{\text{frag}}}, ..., \mathbf{I}_t, ..., \mathbf{I}_{t+l_{\text{frag}}}\}$. We fix $l_{\text{frag}} = 10$ for all our experiments, determined by cross-validation on the iLIDS-VID dataset. Finally, we build a candidate pool (set) of video fragments $G = \{g\}$ by pooling all

the fragments from the sequence Q. It is worth pointing out that some of the obtained fragments of each image sequence can have similar phases of a walking cycle since the local minimum/maximum landmarks of the FEP signal are likely to correspond to certain characteristic walking postures (Figure 6.2). This increases the possibility of finding temporally aligned video fragment pairs (i.e., centred at similar walking postures) given a pair of video fragment sets (G^a, G^b) from two disjoint camera views, facilitating discriminative video fragments selection and matching during model learning. It is observed from Figure 6.2 that the FEP signal may be sensitive to random occlusions and background clutters and lead to noisy/non-characterised video fragments. This usually results in more redundant video fragmentation to some degree. However, it has limited influence on the effectiveness of the learned model, as the proposed selection-and-ranking model in Section 6.1.3 can automatically identify and select discriminative/informative video fragments to train the re-identification ranking model.

Video fragment representation To encode both the dynamics and static appearance information of the subjects, we represent video fragments with both space-time features and colour features. These two types of features complement each other, especially in the context of person re-identification. Colour features have been shown to be significant for person re-identification (Hirzer et al., 2012; Gong et al., 2014a; Liu et al., 2012, 2014a; Zhao et al., 2013a), implicitly capturing the chrome patterns of clothing/appearance that is independent from space-time characteristics of a person's appearance, such as the way people walk. In contrast, the latter is encoded by the space-time features.

Space-time feature Particularly, we exploit HOG3D (Klaser and Marszalek, 2008) as space-time feature representation of a video fragment, due to its advantages demonstrated for applications in action and activity recognition (Wang et al., 2009; Klaser and Marszalek, 2008). In order to capture spatially more detailed and localised space-time information of a person in motion, we decompose a video fragment spatially into 2×5 even cells according to human biological body topology such as head, torso, arms and legs. To capture separately the information of sub-intervals before and after the characteristic walking posture (Figure 6.2 (d)) potentially situated in the middle of a video fragment, the fragment is further divided temporally into two smaller sub-phases, resulting in a total of $20 = 2 \times 5 \times 2$ cells for every single video fragment. Two adjacent cells have 50% overlap for increased robustness to possible spatio-temporal fragment misalignment. A space-time gradient histogram is computed in each cell and then concatenated

to form the HOG3D space-time descriptor $\mathbf{x}_{\text{st}}$ of the fragment g.

Colour feature We adopt the localised average colour histogram as the appearance feature of a video fragment because of its simplicity and effectiveness (Hirzer et al., 2012). Specifically, for each component frame in a video fragment, the colour features are extracted from rectangular patches (16×8 pixels in size) sampled from each frame with an overlap of 8 and 4 pixels vertically and horizontally between each patch (i.e. 50% overlap between adjacent patches). In each patch, we compute the mean values of the HSV and LAB colour channels and form a framewise colour feature vector by concatenating the mean values of all the patches in a frame. To minimise noise and obtain a more reliable colour representation, all the framewise colour features of a fragment are averaged over time to produce a fragment-wise appearance representation $\mathbf{x}_{\text{a}}$ of that fragment g.

Finally, both space-time and colour appearance features $\mathbf{x}_{\text{st}}$ and $\mathbf{x}_{\text{a}}$ are concatenated into a fragment descriptor $\mathbf{x} = [\mathbf{x}_{\text{st}}; \mathbf{x}_{\text{a}}]$. Note, the image frames of all sequences are normalised into a fixed size (128×64 pixels in our implementation) before computing any features.

Notations Formally, for the k-th fragment $g_{i,k}^a$ from the person p_i's image sequence captured in camera a, its descriptor is denoted by $\mathbf{x}_{i,k}^a$. The same is for $g_{i,k}^b$ and $\mathbf{x}_{i,k}^b$. We denote $X_i^a = \{\mathbf{x}_{i,k}^a\}_{k=1}^{|X_i^a|}$ and $X_i^b = \{\mathbf{x}_{i,k}^b\}_{k=1}^{|X_i^b|}$ as the descriptor set for the fragments segmented from the image sequences Q_i^a and Q_i^b of person p_i in camera a and b respectively, where $|\cdot|$ represents the cardinality of a set. The entire collection of descriptors for n training person image sequence pairs $\{(Q_i^a, Q_i^b)\}_{i=1}^n$ (Section 6.1.1) is denoted as $\{(X_i^a, X_i^b)\}_{i=1}^n$.

6.1.3 Selection and Ranking

As shown in Figure 6.2, the fragments of a person image sequence can be contaminated by unknown occlusions and background dynamics, and may also be extracted at an arbitrary time-instance of a walking phase. Given such noisy fragment pair collections generated from cross-view person image sequences, a significant challenge for sequence matching based re-identification is how to identify and select discriminative/informative and temporally aligned fragment pairs (rather than the entire sequences) to learn a suitable ranking model. Formally, the objective is to learn a linear ranking function on the entry-wise absolute difference of two cross-view fragments $\mathbf{x}^a$ and $\mathbf{x}^b$:

$$h(\mathbf{x}^a, \mathbf{x}^b) = \boldsymbol{w}^\top \text{abs}(\mathbf{x}^a - \mathbf{x}^b). \tag{6.3}$$

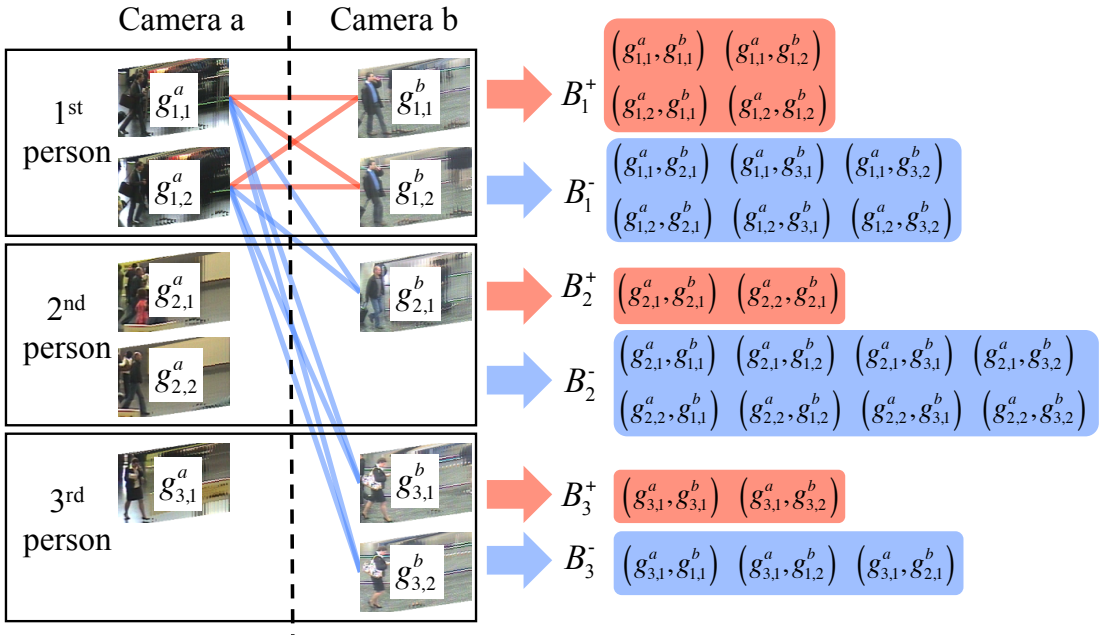


Figure 6.3: The process of constructing the positive and negative bags of fragments for representing individual person sequences (examples of three people are shown). In this instance, symbol $g^a_{1,2}$ refers to the second fragment from the first person image sequence captured in camera a, and similar for others. We form separately a positive (B_i^+) and negative (B_i^-) bag for the i-th person. Take the first person as an example, the cross-view pairings (red lines) of the fragments from the first person form its positive bag B_1^+, while those pairings (blue lines) across the first person and any different person are used to create the negative bag B_1^-.

We assume that for each person there exists at least one cross-view fragment pair that is sufficiently aligned over time and carries desired identity-sensitive information for this person. Our aim is to construct a model capable of automatically discovering and locating not only the best cross-view fragment pair but also multiple cross-view fragment pairs that are sufficiently aligned and discriminative for re-identification. For model training with the best fragment pair, it is equivalent to constraining a ranking function h to prefer the most discriminative cross-view fragment pair of the same person p_i over the pairings over p_i and any other person p_j, $i \neq j$, i.e.

$$\left(\max_{\mathbf{x}^a_{i,\cdot} \in X^a_i, \mathbf{x}^b_{i,\cdot} \in X^b_i} h(\mathbf{x}^a_{i,\cdot}, \mathbf{x}^b_{i,\cdot}) \right) > h(\mathbf{x}^a_{i,\cdot}, \mathbf{x}^b_{j,\cdot}), \forall\, j \neq i, \tag{6.4}$$

For notation simplicity, we define $\boldsymbol{z}^+_{i,k} = \text{abs}(\mathbf{x}^a_{i,\cdot} - \mathbf{x}^b_{i,\cdot})$ as the k-th *positive instance* of person p_i, i.e., the entry-wise absolute difference of two cross-view fragments of the same person p_i, and $\boldsymbol{z}^-_{i,k} = \text{abs}(\mathbf{x}^a_{i,\cdot} - \mathbf{x}^b_{j,\cdot}), j \neq i$ as the k-th *negative instance*, i.e., the absolute difference of two cross-view fragments of p_i and another person. For each person p_i, we form a *positive bag* $B_i^+ = \{\boldsymbol{z}^+_{i,k}\}_{k=1}^{|B_i^+|}$ by pooling the positive instances, and a *negative bag* $B_i^- = \{\boldsymbol{z}^-_{i,k}\}_{k=1}^{|B_i^-|}$ by pooling the negative instances. Figure 6.3 illustrates the formation process of positive and negative bags

for individual persons. By redefining the ranking function $h(\mathbf{x}^a, \mathbf{x}^b) = \tilde{h}(\text{abs}(\mathbf{x}^a - \mathbf{x}^b)) = \tilde{h}(\boldsymbol{z})$, Equation (6.4) can be rewritten as

$$\left(\max_{\boldsymbol{z}_{i,\cdot}^+ \in B_i^+} \tilde{h}(\boldsymbol{z}_{i,\cdot}^+) \right) > \tilde{h}(\boldsymbol{z}_{i,\cdot}^-), \forall \boldsymbol{z}_{i,\cdot}^- \in B_i^- . \tag{6.5}$$

With the ranking constraints in Equation (6.5), we aim to automatically discover and select the most discriminative/informative and temporally aligned cross-view fragment pair $\boldsymbol{z}_{i,\cdot}^+$ within the positive bag B_i^+ for each person p_i for learning an identity discriminative ranking model. To that end, we introduce a binary selection variable $\boldsymbol{v}_i$ with each entry being either 0 or 1 and of unity l_0 norm for each person p_i, and then obtain

$$\tilde{h}(\boldsymbol{Z}_i \boldsymbol{v}_i) > \tilde{h}(\boldsymbol{z}_{i,\cdot}^-), \forall \boldsymbol{z}_{i,\cdot}^- \in B_i^-, \tag{6.6}$$

where each column of $\boldsymbol{Z}_i$ corresponds to one $\boldsymbol{z}^+ \in B_i^+$, and $||\boldsymbol{v}_i||_0 = 1$, $\boldsymbol{e}^\top \boldsymbol{v}_i = 1$, $\boldsymbol{e}$ denotes a vector with each elementary being 1.

To achieve good generalisation ability for the ranking model given the ranking constraints in Equation (6.6), we formulate our problem as a max-margin ranking problem by defining the objective function as:

$$\begin{aligned} \boldsymbol{w}^* = \arg\min_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{\xi}} & \frac{1}{2}||\boldsymbol{w}||^2 + \eta \boldsymbol{e}^\top \boldsymbol{\xi} \\ \text{s.t. } & \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} - (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w} \geq 1 - \xi_{i,k}, \\ & \xi_{i,k} \geq 0, \ \forall \boldsymbol{z}_{i,k}^- \in B_i^-, \ k \in \{1, \ldots, |B_i^-|\}, \\ & ||\boldsymbol{v}_i||_0 = 1, \ \boldsymbol{e}^\top \boldsymbol{v}_i = 1, \ \ i \in \{1, \ldots, n\}. \end{aligned} \tag{6.7}$$

where $\boldsymbol{w}$ is the parameter of the objective ranking function as defined in Equation (6.3), and n the number of persons in the training set. $\boldsymbol{v}$ is the concatenation of the binary selection variables of all persons: $\boldsymbol{v} = [\boldsymbol{v}_1; \boldsymbol{v}_2; ... \boldsymbol{v}_n]$. $\boldsymbol{\xi}$ is the flattened slack variable, formed by all the possible $\xi_{i,k}$. We solve Equation (6.7) by iteratively optimising $\boldsymbol{w}$ and $\boldsymbol{v}$ between a ranking step and a selecting step. η is a balancing parameter, which is set by cross-validation.

Ranking step We fix $\boldsymbol{v}$ to optimise $\boldsymbol{w}$. Equation (6.7) turns into

$$\begin{aligned} \boldsymbol{w}^* = \arg\min_{\boldsymbol{w},\boldsymbol{\xi}} & \frac{1}{2}||\boldsymbol{w}||^2 + \eta \boldsymbol{e}^\top \boldsymbol{\xi} \\ \text{s.t. } & \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} - (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w} \geq 1 - \xi_{i,k}, \\ & \xi_{i,k} \geq 0, \ \forall \boldsymbol{z}_{i,k}^- \in B_i^-, \ k \in \{1, \ldots, |B_i^-|\}, \\ & i \in \{1, \ldots, n\}. \end{aligned} \tag{6.8}$$

With the fragment selections $\boldsymbol{v}$ known, Equation (6.8) is a standard RankSVM problem and can be efficiently solved with a primal training algorithm (Chapelle and Keerthi, 2010).

Selecting step We fix $\boldsymbol{w}$ to optimize $\boldsymbol{v}$. The term on $\boldsymbol{w}$ (i.e. $\frac{1}{2}||\boldsymbol{w}||^2$) can be eliminated and Equation (6.7) becomes

$$\begin{aligned}
&\boldsymbol{v}^* = \arg\min_{\boldsymbol{v},\boldsymbol{\xi}} \boldsymbol{e}^\top \boldsymbol{\xi} \\
&\text{s.t. } \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} - (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w} \geq 1 - \xi_{i,k}, \\
&\qquad \xi_{i,k} \geq 0,\ \forall \boldsymbol{z}_{i,k}^- \in B_i^-,\ k \in \{1,\ldots,|B_i^-|\} \\
&\qquad ||\boldsymbol{v}_i||_0 = 1,\ \boldsymbol{e}^\top \boldsymbol{v}_i = 1,\ \ i \in \{1,\ldots,n\}.
\end{aligned} \tag{6.9}$$

Considering that the person-wise $\boldsymbol{v}_i$ is associated only with $\{\xi_{i,k}\}_{k=1}^{|B_i^-|}$ and we are optimising the summation of all possible $\xi_{i,k}$, Equation (6.9) is equivalent to optimising $\boldsymbol{v}_i$ for each person p_i separately, as

$$\begin{aligned}
&\boldsymbol{v}_i^* = \arg\min_{\boldsymbol{v}_i,\boldsymbol{\xi}_i} \boldsymbol{e}^\top \boldsymbol{\xi}_i \\
&\text{s.t. } \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} - (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w} \geq 1 - \xi_{i,k}, \\
&\qquad \xi_{i,k} \geq 0,\ \forall \boldsymbol{z}_{i,k}^- \in B_i^-,\ k \in \{1,\ldots,|B_i^-|\} \\
&\qquad ||\boldsymbol{v}_i||_0 = 1,\ \boldsymbol{e}^\top \boldsymbol{v}_i = 1.
\end{aligned} \tag{6.10}$$

where $\boldsymbol{\xi}_i = [\xi_{i,1},\ldots,\xi_{i,|B_i^-|}]^\top$. The inequality constraints in Equation (6.10) can be transformed as

$$\begin{aligned}
&\xi_{i,k} \geq 1 - \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} + (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w}, \\
&\xi_{i,k} \geq 0.
\end{aligned} \tag{6.11}$$

Therefore, for any particular $\boldsymbol{v}_i \in V$ that holds $||\boldsymbol{v}_i||_0 = 1$ and $\boldsymbol{e}^\top \boldsymbol{v}_i = 1$ in the selecting space V, the entries $\xi_{i,k}^*$ of the optimal $\boldsymbol{\xi}_i^*$ that minimises the summation $\boldsymbol{e}^\top \boldsymbol{\xi}_i$ shall be

$$\xi_{i,k}^* = \max\{0, 1 - \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} + (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w}\}. \tag{6.12}$$

It is obvious that the summation $\boldsymbol{e}^\top \boldsymbol{\xi}_i$ is a function of $\boldsymbol{v}_i$,

$$\begin{aligned}
q(\boldsymbol{v}_i) &= \sum_{k=1}^{|B_i^-|} \xi_{i,k}^* \\
&= \sum_{k=1}^{|B_i^-|} \max\{0, 1 - \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} + (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w}\}.
\end{aligned} \tag{6.13}$$

Finally we can obtain the $\boldsymbol{v}_i^*$ by optimising $q(\boldsymbol{v}_i)$ via:

$$\begin{aligned}\boldsymbol{v}_i^* &= \arg\min_{\boldsymbol{v}_i \in V} q(\boldsymbol{v}_i) \\ &= \arg\min_{\boldsymbol{v}_i \in V} \sum_{k=1}^{|B_i^-|} \max\{0, 1 - \boldsymbol{v}_i^\top \boldsymbol{Z}_i^\top \boldsymbol{w} + (\boldsymbol{z}_{i,k}^-)^\top \boldsymbol{w}\}, \\ &\text{s.t. } ||\boldsymbol{v}_i||_0 = 1,\ \boldsymbol{e}^\top \boldsymbol{v}_i = 1.\end{aligned} \tag{6.14}$$

For each person p_i, we only have a limited number of $\boldsymbol{v}_i$ in V. Therefore Equation (6.14) can be efficiently solved even with a greedy search.

To begin the model training process, we set $\boldsymbol{v}_i = \frac{1}{|B_i^+|}\boldsymbol{e}$ to initiate a balanced/moderate start since the quality of $\boldsymbol{z}_{i,\cdot}^+$ is unknown *a priori*. The iteration terminates when $\boldsymbol{v}_i$ does not change any more. Typically, the training process stops after $4 \sim 5$ iterations. For learning efficiency, 10% out of all the $\boldsymbol{z}_{i,\cdot}^-$ are randomly selected to form B_i^-. Since only a single $\boldsymbol{z}_{i,\cdot}^+$ for each person p_i is selected and utilised for model learning, we call this model **DVR(single)**.

Multiple Cross-View Fragment Pairs Selection

Thus far we have detailed the procedure of training our DVR (single) model via identifying the *best* cross-view fragment pair in each positive bag B_i^+ (corresponding to a particular person) for learning the ranking function (Equation (6.3)). This allows us to largely avoid the contamination effect from harmful data. Nevertheless, we may simultaneously lose some useful information from discarding the majority of instances $\boldsymbol{z}_{i,\cdot}^+$ of each bag B_i^+, as some of these ignored $\boldsymbol{z}_{i,\cdot}^+$ can be of good-quality. To identify and exploit these good though not-best fragment data $\boldsymbol{z}_{i,\cdot}^+$ is likely to benefit the model learning. To that end, we shall describe next our multiple cross-view fragment pairs selection algorithm for achieving better harness of image sequence data.

Our multiple fragment-pair selection algorithm is based on a goodness/quality measure of individual $\boldsymbol{z}_{i,\cdot}^+$. Once all instances $\boldsymbol{z}_{i,\cdot}^+$ of person p_i are measured by assigning a score $\gamma_{i,\cdot}$ (higher is better) to each instance, we can easily locate multiple (top-$\tilde{k}$) discriminative $\boldsymbol{z}_{i,\cdot}^+$ from the ranked list of all $\boldsymbol{z}_{i,\cdot}^+$ sorted in descending order by the score $\gamma_{i,\cdot}$. Formally, we define $\gamma_{i,\cdot}$ for each $\boldsymbol{z}_{i,\cdot}^+$ as

$$\gamma_{i,\cdot} = \sum_{k=1}^{|B_i^-|} (1 - \xi_{i,k}^*). \tag{6.15}$$

We denote $1 - \xi_{i,k}^*$ as the ranking margin of $\boldsymbol{z}_{i,\cdot}^+$ against the negative instance $\boldsymbol{z}_{i,k}^-$, which can be obtained by Equation (6.12). Given Equation (6.15), the $\boldsymbol{z}_{i,\cdot}^+$ with larger cumulated ranking margin over all the negative instance $\boldsymbol{z}_{i,k}^-$ is preferred. This formulation generalises the single selection

case that searches for the best $\boldsymbol{v}_i^*$ (Equation (6.14)), i.e. the $\boldsymbol{v}_i^*$ and the highest $\gamma_{i,\cdot}$ corresponds to the same selection of positive instance $\boldsymbol{z}_{i,\cdot}^+$.

After the top-$\tilde{k}$ $\boldsymbol{z}_{i,\cdot}^+$ for each person p_i are found and selected, we can obtain multiple (i.e. $\tilde{k}$) $\boldsymbol{v}_i^*$s by setting the corresponding entry of each $\boldsymbol{v}_i^*$ to 1 whilst the remains to 0. We call this model **DVR(top-$\tilde{k}$)**. Similar to the single selection model DVR(single), these ranking constraints associated with the selected top-$\tilde{k}$ $\boldsymbol{z}_{i,\cdot}^+$ are then employed for optimising $\boldsymbol{w}$ with Equation (6.8). In Section 6.3.1, we shall evaluate the effect of different top-$\tilde{k}$ positive instances on the person re-identification performance. An overview of learning the proposed DVR model is presented in Algorithm 4.

DVR Model Complexity

We analyse the training complexity of the DVR model, focusing on the ranking and selecting steps. For our model training, we adopt the primal RankSVM scheme (Chapelle and Keerthi, 2010) as the ranking solver whose complexity is $O(n_{\text{rank}} \times d^2) + O(d^3)$, taken by Hessian computation and the linear search in Newton direction respectively, with n_{rank} and d referring respectively to the number of ranking constraints (see Equations (6.4) and (6.8)) and the feature dimensions. Suppose $\tilde{\eta}$ positive instances per person are selected for learning the ranking function, then $c_{\text{rank}} = \tilde{\eta} \sum_{i=1}^{n} |B_i^-|$, where n is the number of all training people.

The cost for the selection process mainly involves measuring the quality score of each positive instance of all training people with Equation (6.12) and Equation (6.15). Its complexity is $O(c_{\text{rank}} \times d \times u_{\text{rank}})$, where $u_{\text{rank}} = \sum_{i=1}^{n} |B_i^+|$ denotes the total number of positive instance across all training data. The overall complexity of model training including both the ranking and selection steps is $O(c_{\text{rank}} \times d^2 + d^3 + c_{\text{rank}} \times d \times u_{\text{rank}})$. We evaluate and report the model training cost in our experiments (Section 6.3.1).

6.1.4 Person Identity Structure Discovery by DVR

Once learned, the ranking model (Equation (6.3)) can be deployed to perform person re-identification by matching a probe person image sequence Q^{prob} observed in one camera view against a gallery set $\{Q^{\text{gal}}\}$ in another disjoint camera view. Formally, the ranking/matching score of a gallery person sequence Q^{gal} with respect to the probe Q^{prob} is computed as

$$f(Q^{\text{prob}}, Q^{\text{gal}}) = \max_{\mathbf{x}_{i,\cdot} \in X^{\text{prob}}, \mathbf{x}_{j,\cdot} \in X^{\text{gal}}} \boldsymbol{w}^\top \text{abs}(\mathbf{x}_{i,\cdot} - \mathbf{x}_{j,\cdot}), \tag{6.16}$$

Algorithm 4: DVR Learning

Input: Training image sequence pairs $\{(Q_i^a, Q_i^b)\}_{i=1}^n$;

Output: The ranking function $\boldsymbol{w}$ (Equation (6.3));

(I) Video fragmentation (Section 6.1.2):

- Segment each Q into a set of fragments $\{g\}$;
- Extract space-time and appearance features $\mathbf{x}$ from g;

(II) Bag construction (Figure 6.3): for each person p_i,

- Form a positive bag B_i^+ with positive instance $\boldsymbol{z}_{i,\cdot}^+$;
- Form a negative bag B_i^- with negative instance $\boldsymbol{z}_{i,\cdot}^-$;

(III) Learning (Section 6.1.3):

/ Initialise selection vectors */*:

$\boldsymbol{v_i} = \frac{1}{|B_i^+|}, i = 1, ..., n$;

while true **do**

  / Ranking step */*:

  Obtain $\boldsymbol{w}^*$ with fixed $\{\boldsymbol{v}_i\}$ (Equation (6.8));

  / Selecting step */*:

  for $i = 1, ..., n$ **do**

    if single selection **then**

      Obtain $\boldsymbol{v}_i^*$ (Equation (6.14));

    end

    else

      / Multiple selection */*:

      Compute $\gamma_{i,\cdot}$ for each $\boldsymbol{z}_{i,\cdot}^+$ (Equation (6.15));

      Rank $\boldsymbol{z}_{i,\cdot}^+$ with $\gamma_{i,\cdot}$ descendantly;

      Find the top-$\tilde{k}$ $\boldsymbol{z}_{i,\cdot}^+$;

      Obtain $\tilde{k}$ $\boldsymbol{v}_i^*$s (Section 6.1.3);

    end

  end

  / Convergence check */*:

  if no $\boldsymbol{v}_i$ changed **then**

    Return $\boldsymbol{w}^*$.

  end

end

(a) iLIDS-VID

(b) PRID 2011

Figure 6.4: Example pairs of image sequences of the same people appearing in different camera views from (a) the iLIDS-VID dataset, (b) the PRID 2011 dataset. Only every 3rd frame is shown and the total number of frames for each sequence is not identical.

where X^{prob} and X^{gal} are the feature sets of the video fragments extracted from the sequences Q^{prob} and Q^{gal}, respectively. The same video fragmentation process as used for model training (Section 6.1.2) is employed for deploying a trained model. Finally, the gallery persons are sorted in descending order of their assigned matching scores to generate a ranking list.

Combination with prior spatial feature based models Our approach can complement existing spatial feature based re-identification approaches. In particular, we incorporate Equation (6.16) into the ranking scores $\mathcal{R}_i$ obtained by other models as

$$\hat{f}(Q^{\text{prob}}, Q^{\text{gal}}) = \sum_i \alpha_i \mathcal{R}_i(Q^{\text{prob}}, Q^{\text{gal}}) + f(Q^{\text{prob}}, Q^{\text{gal}}), \tag{6.17}$$

where α_i refers to the weighting assigned to the i-th method, which is estimated by cross-validation.

6.1.5 Discussion on Related Models

We discuss the relationship of the proposed DVR model with other relevant contemporary models in the literature, with a focus on their differences. First, most existing max-margin ranking methods (Chapelle and Keerthi, 2010; Prosser et al., 2010) do not consider uncertainty in the ranking constraints during model optimisation. In contrast, the proposed DVR model jointly optimises both the selection of the ranking constraints and the ranking function. This is necessary because the bag-level (e.g. image sequences) supervision cannot directly determine the instance-level (e.g. fragments) constraints (Section 6.1.3).

Second, our model also differs notably from other multi-instance ranking models (Bergeron et al., 2008, 2012; Hu et al., 2008) in a number of aspects as follows: (1) Bergeron et al. (2008) relaxed the selection vectors $\boldsymbol{v}_i$ (Equation (6.6)) to be continues during model optimisation, whilst our model searches for exact solutions of instance selection. As shown in our evaluation (Section 6.3.1), Bergeron et al.'s relaxation method can significantly increase the cost of constraint selection when the training set is large, though it does not compromise the model performance. (2) The model presented in (Hu et al., 2008) focuses on encoding bag-level (or sample-level) constraints into the ranking function by modelling instance-level constraints, assuming all instances can provide contributions to model optimisation. In contrast, we emphasise the selection of discriminative/informative instance data (e.g. fragments) for robust learning, necessary for coping with very noisy and incomplete data (e.g. unregulated image sequences), whilst the stronger assumption made in (Hu et al., 2008) is less valid therein. (3) Different from all these multi-instance

models (Bergeron et al., 2008, 2012; Hu et al., 2008), the proposed DVR model is unique in its capability for allowing different quantities of explicit discriminative instance selection and then exploitation, due to our formulation of a principled instance quality measure (Equation (6.15)). This can potentially increase the flexibility and scalability of our model in a variety of problem settings (e.g. varying degrees of noise) and applications (e.g. other sequence matching based tasks).

6.2 Datasets and Experimental Settings

Datasets Extensive experiments were conducted on two image sequence datasets designed for person re-identification, the PRID 2011 dataset (Hirzer et al., 2011) and the iLIDS Video re-IDentification (iLIDS-VID) dataset (Wang et al., 2014b).

The iLIDS-VID dataset Our new iLIDS-VID person sequence dataset (Wang et al., 2014b) was created based on two non-overlapping camera views from the i-LIDS Multiple-Camera Tracking Scenario (MCTS) (UK , 2008), which was captured at an airport arrival hall under a multi-camera CCTV network. It consists of 600 image sequences for 300 randomly sampled people, with one pair of image sequences from two disjoint camera views for each person. Each image sequence has a variable length consisting of 23 to 192 image frames, with an average number of 73. This dataset is very challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions (Figure 1.6 and Figure 6.4 (a)).

The PRID 2011 dataset The PRID 2011 re-identification dataset (Hirzer et al., 2011) includes 400 image sequences for 200 people from two camera views that are adjacent to each other. Each image sequence has a variable length consisting of 5 to 675 image frames [1], with an average number of 100. Compared with the iLIDS-VID dataset, it is less challenging due to captured in uncrowded outdoor scenes with relatively simple and clean background and rare occlusions (Figure 6.4 (b)).

Evaluation metrics Person ReID results are shown in Cumulated Matching Characteristics (CMC) curves. To obtain stable statistical results, we repeat the experiments for 10 trials and report the average results.

[1] Sequences with more than 21 frames from 178 persons are used in our experiments.

Table 6.1: Comparing different variants of the proposed DVR model in top matching rates (%). Video fragments are represented with HOG3D&Colour. (RT: Ranking Time; ST: Selecting Time; Unit is second).

Dataset	PRID 2011						iLIDS-VID					
Rank *R* (%)	*R*=1	*R*=5	*R*=10	*R*=20	RT	ST	*R*=1	*R*=5	*R*=10	*R*=20	RT	ST
DVR(float)	38.9	68.8	81.1	91.3	**6**	**8**	36.8	59.3	70.9	80.1	**28**	740
DVR(single)	38.9	68.8	81.1	91.3	**6**	9	36.8	59.3	70.9	80.1	**28**	**97**
DVR(top-2)	39.4	70.6	83.7	91.8	13	9	37.7	60.1	71.1	81.4	42	**97**
DVR(top-3)	40.0	**71.7**	84.5	92.2	15	9	39.5	61.1	71.7	81.0	58	**97**
DVR(top-4)	40.0	71.6	84.0	92.8	20	9	39.2	**62.3**	71.7	**81.9**	70	**97**
DVR(top-5)	**40.8**	**71.7**	**84.9**	**93.1**	21	9	**39.9**	62.1	**71.9**	**81.9**	81	**97**

Implementation details From both datasets, the total pool of sequence pairs is randomly split into two subsets of equal size, one for training and one for testing. Following the evaluation protocol on the PRID 2011 dataset (Hirzer et al., 2011), in the testing phase, the sequences from the first camera are used as the probe set while the ones from the other camera as the gallery set.

6.3 Experiments and Evaluations

6.3.1 Evaluation on Model Variants

We evaluated and analysed the proposed DVR model in two perspectives: (1) the effectiveness of different selection mechanisms; (2) the effectiveness of different fragment representations (i.e. HOG3D and HOG3D&Colour) on person re-identification, given a chosen selection algorithm. Note that HOG3D&Colour is adopted as the default fragment representation in our DVR model throughout the following experiments, unless specified otherwise.

For the selection mechanism, we conducted two comparisons: (a) the DVR(single) model *versus* our preliminary model reported in (Wang et al., 2014b) which we call **DVR(float)** since its selection involves a (float) weighted combination of instances in contrast to our new single or multiple explicit instance selection strategies, (b) single *versus* multiple fragment-pairs selection (Section 6.1). The results in Table 6.1 (the first two rows) show that identical scores are obtained by DVR(single) and DVR(float) (Wang et al., 2014b). This is further verified by the observation that both models select almost identical discriminative video fragments. On the other hand, the computational cost and time required are different for these two models, in particular when the content in the data is more crowded therefore selection becomes harder. More specifically, for model training including both the ranking and selecting steps, Table 6.1 shows that both models require similar time for the ranking step of model training on both datasets. This is because

Table 6.2: Comparing different video fragment representation in top matching rates (%) using the DVR (single) model.

Dataset	PRID 2011				iLIDS-VID			
Rank R (%)	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
HOG3D	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.4
Colour	30.1	54.3	64.9	79.7	24.2	44.6	56.0	67.4
HOG3D&Colour	**38.9**	**68.8**	**81.1**	**91.3**	**36.8**	**59.3**	**70.9**	**80.1**

they are subject to the same number of ranking constraints (Equation (6.8)). However, although the time required for the selection routine is similar for the PRID 2011 dataset, DVR(single) is significantly faster that DVR(float) for iLIDS-VID, resulting in over 7$\times$ speed-up. This was performed on a 64-bit Intel CPU Processor @ 2.7 GHz with a MATLAB implementation in Linux OS. These observations suggest that there is no advantage in considering the selector vector being a float weighted combination of instances as originally defined in (Wang et al., 2014b; Bergeron et al., 2008).

One may ask the question how many discriminative fragment pairs should be selected from each cross-view image sequence pair of a person during model training. To that end, we evaluated the performance of re-identification using different numbers of positive fragment pair per person.

It is evident from Table 6.1 that the use of additional discriminative fragment pairs can further boost the overall performance of person re-identification at the price of increased model training time. This empirically supports our analysis on the potential benefits of multiple fragment pair selection and exploitation as discussed in Section 6.1.3. However, the margin of improvement from additional fragment data quickly diminishes. In our experiments, we utilised upto the top rank-5 fragment pairs per person. Any further addition of more pairs had very limited effect in improving the learned ranking model. Moreover, it is also observed that the construction of ranking constraints in RankSVM is a time consuming process and its complexity is linear to the number of constraints. Empirically, selecting the top 3 discriminative fragment pairs from a matched training image sequence pair for model learning provides a good trade-off between re-identification accuracy and model learning cost. For the remaining experiments reported in this section, a DVR(top-3) model was trained for both datasets in the comparative evaluation against other baseline methods.

It is worth pointing out that our preliminary work presented in (Wang et al., 2014b) is somewhat limited on fragment representation as no colour appearance information is considered. Here

Table 6.3: Comparison with gait recognition and temporal sequence matching methods in top matching rates (%).

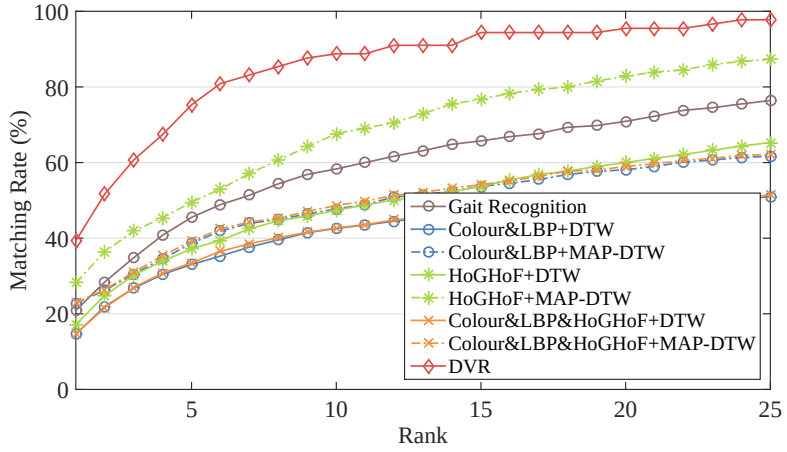
Input	Dataset	PRID 2011				iLIDS-VID			
	Rank R (%)	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Video	Gait Recognition (Martín-Félez and Xiang, 2012)	20.9	45.5	58.3	70.9	2.8	13.1	21.3	34.5
	Colour&LBP+DTW(Rabiner and Juang, 1993)	14.6	33.0	42.6	47.8	9.3	21.7	29.5	43.0
	HoGHoF+DTW	17.2	37.2	47.4	60.0	5.3	16.1	29.7	44.7
	Colour&LBP&HoGHoF+DTW	14.7	33.5	42.7	47.8	10.1	22.5	29.9	43.6
Fragment	Colour&LBP+MAP-DTW	22.8	38.7	47.8	58.1	13.7	28.1	34.9	49.4
	HoGHoF+MAP-DTW	28.2	49.6	67.6	82.8	12.7	34.5	47.2	64.9
	Colour&LBP&HoGHoF+MAP-DTW	22.9	42.5	48.7	59.0	13.9	28.7	36.8	50.8
	DVR (ours)	**40.0**	**71.7**	**84.5**	**92.2**	**39.5**	**61.1**	**71.7**	**81.0**

we report a significant improvement in performance from combining the space-time features (HOG3D) with colour features (Section 3.2). For the DVR(single) model, Table 6.2 shows 34.6% and 57.9% increase in Rank-1 recognition rate on PRID 2011 and iLIDS-VID respectively, as compared to the results reported in (Wang et al., 2014b). This is because the colour information is another important cue for re-identifying people, as indicated by the re-identification performance when only colour features are utilised to represent video fragments. These results demonstrate the importance of utilising both space-time and colour appearance information for person re-identification in video sequence data, further supporting previous studies on the importance of leveraging colour information for re-identification as reported in the literature (Hirzer et al., 2012; Gong et al., 2014a; Liu et al., 2012, 2014a; Zhao et al., 2013a).

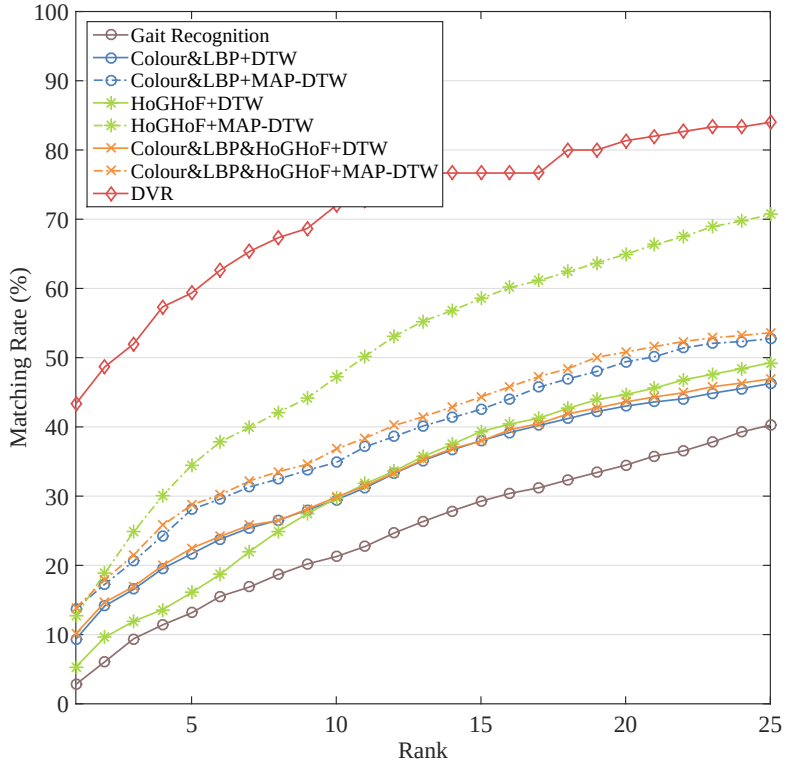
6.3.2 Comparing Gait Recognition and Temporal Sequence Matching

We compared the proposed DVR model with contemporary gait recognition and temporal sequence matching methods for person (re-)identification:

1. Gait recognition (GEI+RSVM) (Martín-Félez and Xiang, 2012): A state-of-the-art gait recognition model using Gait Energy Image (GEI) (Han and Bhanu, 2006) (which is computed from pre-segmented silhouettes in their datasets) as sequence representation and RankSVM (Chapelle and Keerthi, 2010) for recognition. A challenge for applying gait recognition to unregulated person sequences in re-identification scenarios is to generate good gait silhouettes as input. To that end, we first deployed the DPAdaptiveMedianBGS algorithm provided in the BGSLibrary (Sobral, 2013) to extract silhouettes from video sequences in the PRID 2011 and iLIDS-VID datasets, respectively. This approach produces better foreground masking than other alternatives.



(a) PRID 2011



(b) iLIDS-VID

Figure 6.5: Compare CMC curves of the DVR model, gait recognition and temporal sequence matching based methods.

2. DTW (Rabiner and Juang, 1993): We applied Dynamic Time Warping (Rabiner and Juang, 1993) to compute the similarity between two sequences, using either Colour&LBP (Hirzer et al., 2012) or HoGHoF (Laptev et al., 2008) or their combination as per-frame feature descriptor. This is similar to the approach of Simonnet et al. (2012), except that they only used colour features. In comparison, Colour&LBP is a stronger representation as it encodes both colour and texture. Alternatively, HoGHoF encodes both texture and motion information. This baseline method is called in form of "feature+DTW", e.g. Colour&LBP+DTW.
3. MAP-DTW: Besides using the holistic sequences for matching, we also utilised the fragments segmented by our fragmentation algorithm (Section 6.1.2) together with the DTW model for matching person sequences. Specifically, given two sequences, we first obtained their video fragments. Then we performed cross-sequence pairwise fragment matching using the DTW model. The same three types of visual features were exploited as above. Finally, we selected and used the most-matched fragment pair, i.e. the minimal matching distance, to estimate the matching score between the two sequences. This method allows to perform a certain degree of data selection. We call this baseline as "feature+MAP-DTW" in that it selects the most probable fragment pairs as the eventual matching.

Table 6.3 and Figure 6.5 show the comparative results between DVR, GEI+RSVM (gait), DTW using visual features Colour&LBP+DTW, HoGHoF+DTW, Colour&LBP&HoGHoF, and MAP-DTW based on fragment-level matching. It is evident that the proposed DVR outperforms significantly all others on both datasets.

In particular, gait recognition (Martín-Félez and Xiang, 2012) achieves the worst re-identification accuracy on the iLIDS-VID dataset. This is largely due to very noisy GEI features available from person sequences. This is evident from the examples shown in Figure 6.6 : the extracted gait foreground masks tend to be affected/contaminated by other moving objects in the scene, whilst our DVR model trains itself by simultaneously selecting and ranking only those fragments of image sequences which suffer the least from occlusions and noise. Given the uniform background and non-crowded scene in the PRID 2011 dataset, gait recognition obtains reasonably good accuracy on it. Moreover, DTW based sequence matching for re-identification using either Colour&LBP, HoGHoF, or their combination also suffer notably from the inherent uncertain nature of re-identification sequences and perform significantly poorer than the proposed

(a) PRID

(b) iLIDS-VID

Figure 6.6: Two examples of GEI gait features and our video fragment pairs, each in one row. In both examples, the leftmost thumbnail shows GEI gait features, while the remaining thumbnails present some examples of fragment pairs, with the automatically selected pairs marked by red bounding boxes. A fragment is visualised as the weighted average of all its frames with emphasis on its central frame.

DVR approach. This is largely due to: (1) Person sequences have different durations with arbitrary starting/ending frames, also potentially different walking cycles. Therefore, attempts to match holistically entire sequences inevitably suffer from mismatching with erroneous similarity measurement. (2) There is no clear (explicit) mechanism to avoid incompleteness/missing data, typical in busy scenes. (3) Direct sequence matching is less discriminative than learning an inter-camera discriminative mapping function explicitly, which is built into the DVR model by exploring multi-instance (fragment-pair) selection and ranking.

By selecting the most-matched fragment pairs as DVR, MAP-DTW consistently improves the people matching accuracy, particularly when using the HoGHoF feature. This suggests our DVR model design principle that data selection is critical given inherently unaligned and inaccurate person sequences. However, without discriminative learning, it is still largely inferior than the proposed DVR model.

6.3.3 Comparing Spatial Feature Representations

To evaluate the effectiveness of discriminate video fragmentation selection and ranking using both spatial appearance and space-time features for person re-identification, we compared the

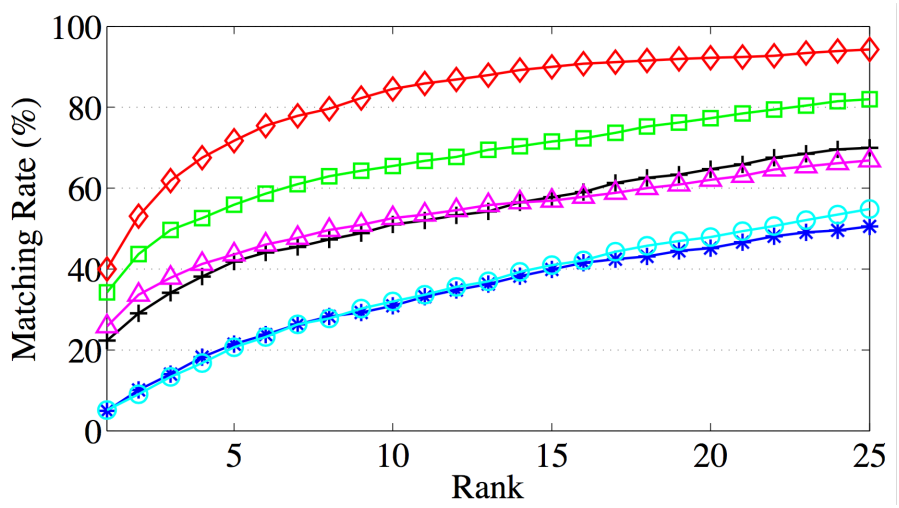
Table 6.4: Comparing spatial appearance feature based re-identification methods in top matching rates (%). (SS: Single-Shot; MS: Multi-Shot).

Dataset	PRID 2011				iLIDS-VID			
Rank *R* (%)	*R*=1	*R*=5	*R*=10	*R*=20	*R*=1	*R*=5	*R*=10	*R*=20
SS-Colour&LBP(Hirzer et al., 2012)+RSVM	22.4	41.8	51.0	64.7	9.1	22.6	33.2	45.5
SS-SDALF (Farenzena et al., 2010)	4.9	21.5	30.9	45.2	5.1	14.9	20.7	31.3
MS-SDALF (Farenzena et al., 2010)	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
Salience (Zhao et al., 2013b)	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
MS-Colour&LBP+RSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
DVR (ours)	**40.0**	**71.7**	**84.5**	**92.2**	**39.5**	**61.1**	**71.7**	**81.0**

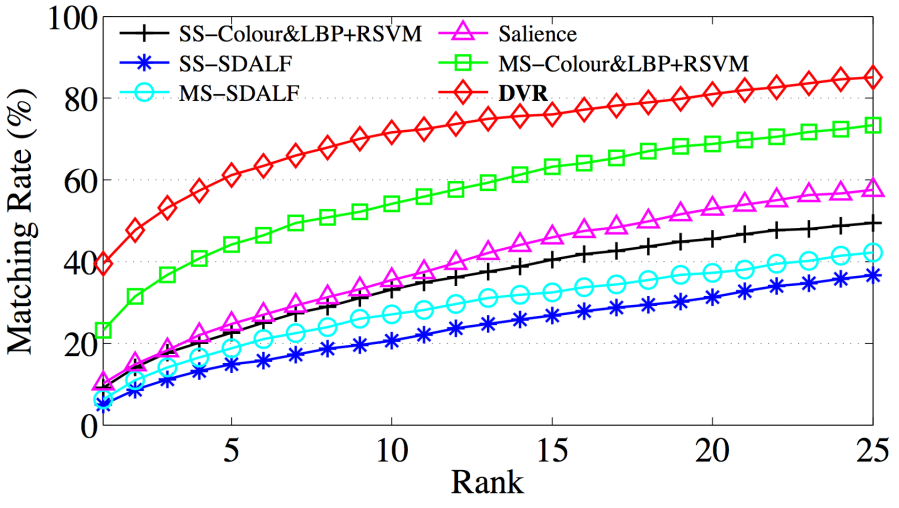
proposed DVR model against a wide range of contemporary re-identification models using spatial features, either in single-shot or as multiple frames (multi-shot). In order to process the iLIDS-VID dataset for our experiments, we mainly considered those methods with both their code available publicly and being contemporary. They include

1. SDALF (Farenzena et al., 2010) (both single-shot and multi-shot versions);
2. Salience (Zhao et al., 2013b);
3. a combination of colour and texture (Colour&LBP) (Hirzer et al., 2012) with RankSVM (Chapelle and Keerthi, 2010) as the distance metric;
4. Moreover, we also extended a Colour&LBP single-shot model to multi-shot by averaging the Colour&LBP features of each frame over a person sequence to focus on stable appearance cues and suppress noise, in a similar approach to (John et al., 2013). We call this method MS-Colour&LBP+RSVM.

Table 6.4 and Figure 6.7 show the results. It is evident that the proposed DVR model outperforms significantly all the spatial feature based methods on both datasets, e.g. 55.0% and 287.3% Rank-1 improvement over Salience, whilst 16.6% and 70.3% Rank-1 improvement over MS-Colour&LBP+RSVM on PRID 2011 and iLIDS-VID, respectively. Note that the improvement margin achieved by the DVR model on iLIDS-VID (a more challenging dataset) is much more significant than that on PRID 2011. This suggests the exceptional effectiveness of the proposed selection based sequence matching in dealing with challenging real-world data for learning a robust re-identification ranking function. More concretely, the power of our DVR model can be largely attributed to identity-sensitive space-time gradient cues learned by our discriminative fragment selection based matching and ranking mechanism, beyond the conventional models of



(a) PRID 2011



(b) iLIDS-VID

Figure 6.7: CMC curve comparison between DVR and existing spatial feature based models. (SS: Single-Shot, MS: Multi-Shot).

only learning from the spatial appearance data, e.g. colour and texture. For a further analysis on the DVR model on its complementarity to existing spatial feature based methods, more details are discussed next.

6.3.4 Complementary to Spatial Features

We further evaluated the complementary effect between the DVR model and existing colour and texture feature based re-identification approaches. The results are reported in Table 6.5. It is evident that for any existing appearance model, significant performance gain was achieved by incorporating the DVR ranking score (Equation (6.17)) into its ranking list. More specifically, the Rank-1 re-identification performance of using multi-shot colour and texture features (MS-Colour&LBP) was boosted by 23.9% and 76.7% on PRID 2011 and iLIDS-VID respectively; Rank-1 by Salience feature improved by 86.8% and 302.0%; Rank-1 score of the combination of Salience and MS-SDALF boosted by 92.4% and 304.9%. Such a performance step-change in improving conventional spatial feature based models is primarily due to the exploration of discriminative space-time features by the proposed DVR selection model. This space-time selective matching process discovers mostly independent source of information as compared to all the static appearance features, therefore playing a significant complementary and beneficial role to contemporary spatial feature based models. It is also worth pointing out that most existing spatial feature based methods benefit more from combining with the DVR model when tested against the iLIDS-VID dataset, and less so on the PRID 2011. This observation highlights the importance and necessity of discriminative fragment selection for robust model learning given video data from more crowded public scenarios where blindly learning a model from all the data without selection leads to poorer and degraded models.

It is also evident from Table 6.5 that the DVR model can benefit from combining with other spatial feature based re-identification models, although slightly. This gain may be explained as the result by drawing diverse sources of spatial features. Overall, the best results are produced by Salience&MS-SDALF+DVR, suggesting effective complementary benefit among them all.

6.3.5 Evaluation on Space-time Fragment Selection

To evaluate the space-time video fragment selection mechanism in the proposed DVR model, we implemented two baseline methods without this selection mechanism:

Table 6.5: The complementary effect of DVR to existing spatial features based models in top matching rates (%). (MS: Multi-Shot).

Dataset	PRID 2011				iLIDS-VID			
Rank *R* (%)	*R*=1	*R*=5	*R*=10	*R*=20	*R*=1	*R*=5	*R*=10	*R*=20
DVR (ours)	40.0	71.7	84.5	92.2	39.5	61.1	71.7	81.0
MS-Colour&LBP+RSVM	34.3	56.0	65.5	77.3	23.2	44.2	54.1	68.8
MS-Colour&LBP+DVR	42.5	70.1	83.5	92.8	41.0	62.1	**73.6**	82.5
MS-SDALF (Farenzena et al., 2010)	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
MS-SDALF+DVR	44.2	71.2	85.1	92.5	40.9	62.7	72.1	82.1
Salience (Zhao et al., 2013b)	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
Salience+DVR	48.2	**75.2**	87.0	94.2	41.0	**63.7**	72.7	**83.3**
Salience&MS-SDALF	25.1	42.9	52.0	62.2	10.2	25.3	35.2	52.9
Salience&MS-SDALF+DVR	**48.3**	74.9	**87.3**	**94.4**	**41.3**	63.5	72.7	83.1

Table 6.6: The effect of space-time video fragment selection. (SS: Single-Shot, MS: Multi-Shot).

Dataset	PRID 2011				iLIDS-VID			
Rank *R* (%)	*R*=1	*R*=5	*R*=10	*R*=20	*R*=1	*R*=5	*R*=10	*R*=20
SS-HOG3D&Colour+RSVM	25.7	49.1	61.0	75.5	15.5	33.7	47.5	61.4
MS-HOG3D&Colour+RSVM	29.6	54.9	70.8	86.1	19.9	42.4	53.6	67.6
DVR (ours)	**40.0**	**71.7**	**84.5**	**92.2**	**39.5**	**61.1**	**71.7**	**81.0**

1. SS-HOG3D&Colour+RSVM: Each person sequence is represented by the HOG3D&Colour descriptor of a single fragment randomly selected from the image sequence;
2. MS-HOG3D&Colour+RSVM: Each person sequence is represented by the averaged HOG3D & Colour descriptor of four fragments uniformly selected from the sequence. In both baseline methods, RankSVM (Chapelle and Keerthi, 2010) is used to rank the person sequence representations. For a fair comparison, the length of these fragments used for both baselines is set the same as those utilised in our DVR model.

The results are presented in Table 6.6. On the PRID 2011 dataset, the DVR model outperforms SS-HOG3D&Colour +RSVM and MS-HOG3D&Colour+RSVM in Rank-1 by 55.6% and 35.1%, respectively. The performance advantage is even greater on the more challenging iLIDS-VID dataset, with 154.8% and 98.5% in Rank-1 improvement. It demonstrates clearly that in the presence of significant noise and given unregulated person image sequences, it is indispensable to automatically select discriminative space-time fragments from raw image sequences in order to construct a more robust model for person re-identification. It is also noted that MS-

HOG3D&Colour+RSVM outperforms SS-HOG3D&Colour+RSVM by suppressing noise using temporal averaging. Although such a straightforward temporal averaging approach can have some benefits over single-shot methods, it loses important discriminative space-time information when applying uniform temporal smoothing.

6.4 Summary

This chapter has presented a Discriminative Video Ranking (DVR) framework for person identity structure discovery (or person re-identification) by video ranking using discriminative space-time and appearance feature selection. Our extensive evaluations show that this model outperforms a wide range of contemporary techniques from gait recognition and temporal sequence matching to state-of-the-art single-shot/multi-shot/multi-frame spatial feature representation based re-identification models. In contrast to existing re-identification approaches that often employ spatial appearance of person alone, the proposed method is capable of capturing more accurately both appearance and space-time information that are discriminative to person re-identification through learning a cross-view multi-instance ranking function. This is made possible by the ability of our model to automatically discover and exploit the most reliable and informative video fragments extracted from inherently incomplete and inaccurate person image sequences captured against cluttered background, without any guarantee on person walking cycles and starting/ending frame alignment. Moreover, the proposed DVR model significantly complements/improves existing spatial appearance features when combined for person re-identification. Extensive comparative evaluations were conducted to validate the advantages of the proposed model over a variety of baseline methods on two challenging image sequence based re-identification datasets, the PRID 2011 and iLIDS-VID benchmarks. More experiments on different data settings are needed to evaluate further the capacity of the proposed discriminative selection and ranking strategy.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis has presented a collection of data structure discovery methods for automated surveillance video content analysis and understanding (Figure 1 in Abstract). In particular, a variety of data cluster analysis settings for discovering the inherent group structures on single-camera data have been investigated and explored. Besides, multi-camera visual data structure analysis is also studied, owing to the natural necessity and requirement of monitoring more expanded public areas in video surveillance. These problems are inherently challenging due to intrinsic visual ambiguities and various noises, or the significant appearance variations across camera views. Specifically,

1. In Chapter 3, an unsupervised visual data structure discovery or clustering framework is presented for sensing the weak and subtle data similarity. This allows to obtain more accurate and meaningful data neighbourhood structures. The resulting data similarity measures in turn significantly benefit graph based clustering algorithms for uncovering hidden data group patterns.
2. In Chapter 4, a semi-supervised visual data structure discovery approach is formulated for taking into account additional prior knowledge expressed in form of pairwise constraints. This model advances unsupervised clustering mechanisms by effectively incorporating sparse high-level relations between data samples. Therefore, this resulting cluster structures mirror more agreement as human. Importantly, this model is characterised with

the ability of tackling invalid pairwise links, which is mostly ignored in the previous works.

3. In Chapter 5, multi-source data structure discovery is investigated. In contrast to the typical single-source setting (as in the above two models), heterogeneous multi-source data presents additional new challenges for computational methods such as the heteroscedasticity and dimensionality discrepancy problem. To that end, a multi-source data joint learning approach is formulated with characteristics of uncovering and exploiting latent correlations between distinct data sources, and of being able to tackle partial or missing non-visual data. This multi-source model can be applied to perform video summarisation by discovering group structures and inferring high-level labels for new videos.
4. In Chapter 6, the multi-camera data structure discovery problem is considered, complementary to the above three structure analysis methods for single camera view data. Relatively, a unique challenge in the multi-camera data setting is the large viewing condition changes across distributed cameras. Instead of looking into group based events as in Chapter 5, this multi-camera method examines and discover person-specific activity structures over more expanded spaces from partially captured visual observation (a.k.a. person re-identification). Different from most contemporary re-identification methods typically relying on static appearance of people, this model uniquely sets out to select and explore identity discriminative space-time dynamic patterns for recognising people among appearance-similar population and discover person-centric distribution structures in crowded public places.

It is demonstrated that these proposed models have potentials and benefits for dealing with other similar tasks in computer vision, machine learning and pattern recognition, although originally proposed and primarily evaluated on surveillance videos and related applications/tasks. Whilst the newly developed algorithms have touched quite a few issues and problems in surveillance video structure analysis, other directions and dimensions are also possibly promising to investigate and explore, as discussed below.

7.2 Future Work

The potential research directions for future work beyond the proposed methods are summarised as follows to end this thesis.

1. (Chapter 3) **Unsupervised visual data structure discovery**: Unsupervised visual data

cluster structure analysis remains an open problem even though large strides have been made recently. With regards to similarity learning based methods as this proposed model, more further research effort is still in need, e.g. one possible dimension is to combine stronger localised similarity criteria in the proposed framework by defining more complex and effective split functions (e.g. multi-layer perceptions (Rota Bulo and Kontschieder, 2014) and discriminative SVM classifiers (Yao et al., 2011)), or alternative training strategies (e.g. minimising a global objective (Schulter et al., 2013b,a), utilising underlying tree structures (Johnson and Zhang, 2014)), Alternatively, other unsupervised data partitioning criteria can be considered (Yu et al., 2011; Criminisi and Shotton, 2012; Pei et al., 2013).

2. (Chapter 4) **Semi-supervised visual data structure discovery**: The consistency criterion between constraints and data feature representations is shown to be a good metric for noisy link detection. But, how to quantify and measure the consistency degree of individual constraints is still a challenging problem and far from being addressed. Specifically, only a proportion of invalid constraints can be correctly filtered and thus more advanced algorithms are required for this issue. Additionally, it is interesting to derive effective approaches to actively selecting data pairs for nominating informative sample-pair query. The aim is to reduce significantly the human labelling amount while still achieve similar clustering results. Despite the already made efforts in (Basu et al., 2004a; Xu et al., 2005; Mallapragada et al., 2008; Wang and Davidson, 2010), noisy constraints measure and detection (Davidson et al., 2006; Ares et al., 2012; Van Craenendonck and Blockeel, 2015), sparse constraint propagation (Lu and Ip, 2010; Lu and Peng, 2013a; Fu and Lu, 2015), it remains valuable to develop a unified model considering simultaneously the problems of sparse/noisy pairwise constraints and active annotation. Algorithmically, the ideas presented in existing semi-supervised random forests (Tang et al., 2013; Leistner et al., 2009) may be useful in developing new more advanced models.

3. (Chapter 5) **Multi-source data structure discovery**: Our MSC-Forest model has demonstrated favourable capabilities of handling heterogeneous noisy data. This indicates its promising potentials to deal with other visual tasks involving intrinsically noisy data. The potential of this proposed random forest variant for processing imperfect data in more generic pattern recognition and data mining applications is interesting to investigate and explore (Wang et al., 2016). On the other hand, the proposed approach requires to learn

a separate model for each particular scene/camera view. This may limit its generality and scalability in real-world applications. Therefore, it is also interesting to integrate existing and/or design new transfer learning (Pan and Yang, 2010) and domain adaptation techniques (Margolis, 2011) upon the proposed multi-source data learning algorithm.

4. (Chapter 6) **Person identity structure discovery**: The initial effort of exploiting space-time information from image sequences for person structure discovery or ReID has shown effective and encouraging by the proposed Discriminative Video Ranking (DVR) model. Two lines of extension work can be considered. One is to develop a multi-instance learning model based on random forests, as the previous chapters. The other concerns the person ReID setting: similar to common supervised learning methods, this model requires a large number of labelled cross-view pairwise people for each camera pair. This requirement can restrict the scalability in large scale application scenarios, e.g. surveillance camera networks with many camera pairs. It is prohibitively expensive to collect sufficient pairwise labels for all possible camera pairs in real-world settings. Hence, it is worth to explore image sequence based ReID methods requiring less or even no cross-view labels e.g. by adopting semi-supervised learning strategies (Zhu, 2005; Chapelle et al., 2006). Another desired objective is to develop a general ReID model which can be applied to as many camera pairs as possible once properly trained. One ideal solution is unsupervised people matching although existing methods (Wang et al., 2014a; Zhao et al., 2013b; Farenzena et al., 2010) are still largely inferior compared to supervised counterparts (Zhao et al., 2014a; Paisitkriangkrai et al., 2015; Ding et al., 2015).

Bibliography

R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

M. R. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In *IEEE International Conference on Computer Vision*, pages 786–793, Barcelona, Spain, November 2011.

Y. Amit and D. Geman. Randomized inquiries about shape: An application to handwritten digit recognition. Technical report, Department of Statistics, The University of Chicago, November 1994.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, July 1997.

R. Araujo. *A Semi-Supervised Approach for Kernel-Based Temporal Clustering*. PhD thesis, University of Waterloo, April 2015.

M. E. Ares, J. Parapar, and Á. Barreiro. An experimental study of constrained clustering effectiveness in presence of erroneous constraints. *Information Processing & Management*, 48(3): 537–551, May 2012.

A. Asuncion and D. Newman. UCI machine learning repository, 2007.

S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 435–440, Boston, Massachusetts, United States, August 2010.

S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Boosted human re-identification using riemannian manifolds. *Image and Vision Computing*, 30(6):443–452, June 2012.

G. Ball and I. HALL DJ. A novel method of data analysis and pattern classification. isodata, a novel method of data analysis and pattern classification. Technical report, 1965.

K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *IEEE International Conference on Computer Vision*, volume 2, pages 408–415, Vancouver, Canada, July 2001.

K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, October 2010.

J. N. Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37(11):2049, November 1979.

S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *International Conference on Machine learning*, Sydney, Australia, July 2002.

S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *SIAM International Conference on Data Mining*, pages 333–344, Florida, United States, April 2004a.

S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 59–68, Seattle, Washington, United States, August 2004b.

S. Basu, I. Davidson, and K. Wagstaff. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

A. Bedagkar-Gala and S. K. Shah. Part-based spatio-temporal model for multi-person re-identification. *Pattern Recognition Letters*, 33(14):1908–1915, December 2012.

A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, April 2014.

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, pages 585–591, Vancouver, British Columbia, Canada, December 2001.

H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Tracking multiple people under global appearance constraints. In *IEEE International Conference on Computer Vision*, pages 137–144, Barcelona, Spain, November 2011.

C. Bergeron, J. Zaretzki, C. Breneman, and K. P. Bennett. Multiple instance ranking. In *International Conference on Machine learning*, pages 48–55, Helsinki, Finland, July 2008.

C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett. Fast bundle algorithm for multiple-instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1068–1079, June 2012.

G. Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13 (1):1063–1095, January 2012.

M. B. Blaschko and C. H. Lampert. Correlational spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, United States, June 2008.

D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, Toronto, Canada, July 2003.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In *International Conference on Machine learning*, pages 55–63, Madison, Wisconsin, United States, July 1998.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the ACM Annual Conference on Computational Learning Theory*, pages 92–100, Madison, Wisconsin, United States, July 1998.

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

L. Breiman. Rf/tools: A class of two-eyed algorithms. In *SIAM Workshop, Statistics Department, UC Berkeley*, 2003.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. Chapman & Hall/CRC Press, 1984.

X. Cai, F. Nie, H. Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, Colorado, United States, June 2011.

R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *International Conference on Machine learning*, pages 161–168, Pittsburgh, Pennsylvania, United States, June 2006.

R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *International Conference on Machine learning*, Helsinki, Finland, July 2008.

O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval Journal*, 13(3):201–215, June 2010.

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

K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen. An adaptive learning method for target tracking across multiple cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, United States, June 2008.

N. Chen, J. Zhu, F. Sun, and E. P. Xing. Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (12):2365–2378, 2012.

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

T. Coleman, J. Saunderson, and A. Wirth. Spectral clustering with inconsistent advice. In *International Conference on Machine learning*, pages 152–159, Helsinki, Finland, July 2008.

D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *International Conference on Image Analysis and Processing*, pages 179–189, Vietri sul Mare, Italy, September 2009.

A. Criminisi and J. Shotton. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, February 2012.

M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, Providence, Rhode Island, United States, June 2012.

I. Davidson, K. Wagstaff, and S. Basu. Measuring constraint-set utility for partitional clustering algorithms. In *European conference on Principle and Practice of Knowledge Discovery in Databases*, pages 115–126, Berlin, Germany, September 2006.

J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine learning*, pages 209–216, Corvallis, Oregon, United States, June 2007.

J. W. Davis and A. F. Bobick. The representation and recognition of human movement using temporal templates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–934, San Juan, Puerto Rico, United States, June 1997.

D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4): 364–366, 1977.

M. Der and L. Saul. Latent coincidence analysis: A hidden variable model for distance metric learning. In *Advances in Neural Information Processing Systems*, pages 3239–3247, Lake Tahoe, Nevada, United States, December 2012.

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

P. Donmez and J. G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *ACM Conference on Information and Knowledge Management*, pages 619–628, Napa, California, United States, October 2008.

T. D'Orazio, P. Mazzeo, and P. Spagnolo. Color brightness transfer function evaluation for non overlapping multi camera tracking. In *ACM International Conference on Distributed Smart Cameras*, pages 1–6, 2009.

A. Doumanoglou, T.-K. Kim, X. Zhao, and S. Malassiotis. Active random forests: An application to autonomous unfolding of clothes. In *European Conference on Computer Vision*, pages 644–658. Zurich, Switzerland, September 2014.

J. Du and C. X. Ling. Active learning with human-like noisy oracle. In *IEEE International Conference on Data Mining*, pages 797–802, Sydney, Australia, December 2010.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

R. Duin and M. Loog. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 732–739, 2004.

P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, pages 97–112. Copenhagen, Danmark, May 2002.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.

S. R. Fanello, C. Keskin, P. Kohli, S. Izadi, J. Shotton, A. Criminisi, U. Pattacini, and T. Paek. Filter forests for learning data-dependent convolutional kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1709–1716, Columbus, Ohio, United States, June 2014.

M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, San Francisco, California, United States, June 2010.

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

S. Feng, Z. Lei, D. Yi, and S. Z. Li. Online content-aware video condensation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, United States, June 2012.

D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino. Semi-supervised multi-feature learning for person re-identification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 111–116, Krakw, Poland, August 2013.

I. E. Frank and R. Todeschini. *The data analysis handbook*. Elsevier, 1994.

A. Freund, D. Pelleg, and Y. Richter. Clustering from constraint graphs. In *SIAM International Conference on Data Mining*, pages 301–312, Atlanta, Georgia, United States, April 2008.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315 (5814):972–976, February 2007.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):303–316, Feb 2014.

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

Z. Fu and Z. Lu. Pairwise constraint propagation: A survey. *arXiv e-prints*, February 2015.

Z. Fu, H. H. Ip, H. Lu, and Z. Lu. Multi-modal constraint propagation for heterogeneous image clustering. In *ACM International Conference on Multimedia*, pages 143–152, Scottsdale, Arizona, United States, November 2011.

Z. Fu, H. Lu, H. H. Ip, and Z. Lu. Modalities consensus for multi-modal constraint propagation. In *ACM International Conference on Multimedia*, pages 773–776, Nara, Japan, October 2012.

J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, April 2011.

U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A string of feature graphs model for recognition of complex activities in natural videos. In *IEEE International Conference on Computer Vision*, pages 2595–2602, Barcelona, Spain, November 2011.

N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, New York City, New York, United States, June 2006.

S. Gong and T. Xiang. *Visual analysis of behaviour: from pixels to semantics*. Springer, 2011.

S. Gong, C. C. Loy, and T. Xiang. Security and surveillance. In *Visual Analysis of Humans: Looking at People*, pages 455–472. September 2011.

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

Y. Gong. Summarizing audiovisual contents of a video program. *EURASIP Journal on Advances in Signal Processing*, pages 160–169, February 2003.

Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, January 2014b.

J. C. Gower and G. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18(1):54–64, 1969.

D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, pages 262–275, Marseille, France, October 2008.

D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, Rio de Janeiro, Brazil, October 2007.

M. W. Green. The appropriate and effective use of security technologies in us schools. a guide for schools and law enforcement agencies. 1999.

M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, pages 498–505, Kyoto, Japan, September 2009.

L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11(9): 1074–1085, September 1992.

O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ACM International Conference on Distributed Smart Cameras*, pages 1–6, 2008.

J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, February 2006.

J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

P. Hansen and M. Delattre. Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association*, 73(362):397–403, June 1978.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, December 2004.

S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *IEEE International Conference on Computer Vision*, pages 263–270, Barcelona, Spain, November 2011.

J. A. Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

D. Heath, S. Kasif, and S. Salzberg. Induction of oblique decision trees. In *International Joint Conference of Artificial Intelligence*, pages 1002–1007, Chambry, France, August 1993.

J. Heer and E. H. Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In *SIAM International Conference on Data Mining Workshop*, pages 51–58, Chicago, Illinois, United States, April 2001.

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

M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, pages 780–793. Florence, Italy, October 2012.

T. K. Ho. Random decision forests. In *IEEE International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, Montreal, Quebec, Canada, August 1995.

T. K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, August 1998.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, December 1936.

J. Hu, J. Lu, and Y.-P. Tan. Deep transfer metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, United States, June 2015.

Y. Hu, M. Li, and N. Yu. Multiple-instance ranking: Learning to rank images for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, United States, June 2008.

H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen. Affinity aggregation for spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, United States, June 2012.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8): 651–666, June 2010.

A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, February 2008.

K. Jeong and C. Jaynes. Object matching in disjoint cameras using a color transfer approach. *Machine Vision and Applications*, 19(5-6):443–455, october 2008.

V. John, G. Englebienne, and B. Krose. Solving person re-identification in non-overlapping camera using efficient gibbs sampling. In *British Machine Vision Conference*, Bristol, United Kingdom, September 2013.

R. Johnson and T. Zhang. Learning nonlinear functions using regularized greedy forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):942–954, 2014.

S. Jones and L. Shao. Unsupervised spectral dual assignment clustering of human actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 604–611, Columbus, Ohio, United States, June 2014.

K. Kamvar, S. Sepandar, K. Klein, D. Dan, M. Manning, and C. Christopher. Spectral learning. In *International Joint Conference of Artificial Intelligence*, pages 561–566, Acapulco, Mexico, August 2003.

A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, Minnesota, United States, June 2007.

H. Kang, X. Chen, Y. Matsushita, and X. Tang. Space-time video montage. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York City, New York, United States, June 2006.

S. Karaman and A. D. Bagdanov. Identity inference: generalizing person re-identification scenarios. In *European Conference on Computer Vision Workshop*, pages 443–452, Florence, Italy, October 2012.

I. Karydis, A. Nanopoulos, H.-H. Gabriel, and M. Spiliopoulou. Tag-aware spectral clustering of music items. In *International Society for Music Information Retrieval Conference*, pages 159–164, October 2009.

V. Khalidov, F. Forbes, and R. Horaud. Conjugate mixture models for clustering multimodal data. *Neurocomputing*, 23(2):517–557, February 2011.

C. Kim and J.-N. Hwang. Object-based video abstraction for video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1128–1138, December 2002.

A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, Florence, Italy, April 2008.

A. Klaser and M. Marszalek. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, Leeds, United Kingdom, September 2008.

B. Knight and A. Johnston. The role of movement in face recognition. *Visual Cognition*, 4(3): 265–273, 1997.

P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. Geof: Geodesic forests for learning coupled predictors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 65–72, Portland, Oregon, United States, June 2013.

P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bulo'. Deep neural decision forests. In *IEEE International Conference on Computer Vision*, Santiago, Chile, December 2015.

M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, Providence, Rhode Island, United States, June 2012.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, Lake Tahoe, Nevada, United States, December 2012.

B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. *Machine Learning*, 74(1):1–22, January 2009.

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

J. Lai and Y. Yi. Key frame extraction based on visual attention model. *Journal of Visual Communication and Image Representation*, 23(1):114–125, January 2012.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, Miami, Florida, United States, June 2009.

T. Lange, M. H. Law, A. K. Jain, and J. M. Buhmann. Learning with constrained and unlabelled data. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 731–738, San Diego, California, United States, June 2005.

I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from

movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, United States, June 2008.

V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2003.

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

R. Layne, T. M. Hospedales, and S. Gong. Domain transfer for person re-identification. In *ACM International Conference on Multimedia Workshop*, pages 25–32, Barcelona, Spain, October 2013.

R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014a.

R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *British Machine Vision Conference*, Nottingham, United Kingdom, September 2014b.

Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, pages 1–18, January 2015.

Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, United States, June 2012.

C. Leistner, A. Saffari, J. Santner, and H. Bischof. Semi-supervised random forests. In *IEEE International Conference on Computer Vision Workshop*, pages 506–513, Kyoto, Japan, September 2009.

A. Li, L. Liu, and S. Yan. Person re-identification by attribute-assisted clothes appearance. In *Person Re-Identification*, pages 119–138. Springer, 2014a.

W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, Portland, Oregon, United States, June 2013a.

W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, Portland, Oregon, United States, June 2013b.

W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, United States, June 2014b.

Z. Li, J. Liu, and X. Tang. Constrained clustering via spectral regularization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 421–428, Miami, Florida, United States, June 2009.

G. Lian, J.-H. Lai, C. Y. Suen, and P. Chen. Matching of tracked pedestrians across disjoint camera views using ci-dlbp. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(7):1087–1099, July 2012.

S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, United States, June 2015.

D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine learning*, pages 296–304, Madison, Wisconsin, United States, July 1998.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, June 2002.

Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *IEEE International Conference on Computer Vision*, pages 444–451, Kyoto, Japan, September 2009.

B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In *ACM Conference on Information and Knowledge Management*, pages 20–29, McLean, Virginia, United States, November 2000.

C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *European Conference on Computer Vision Workshop*, pages 391–401, Florence, Italy, October 2012.

C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013.

C. Liu, S. Gong, and C. C. Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 47(4):1602–1615, April 2014a.

J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, Colorado Springs, Colorado, United States, June 2011.

X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3557, Columbus, Ohio, United States, June 2014b.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, March 1982.

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

C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995, Miami, Florida, United States, June 2009.

C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing*, Melbourne, Australia, September 2013.

Z. Lu and M. A. Carreira-Perpinán. Constrained spectral clustering through affinity propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, United States, June 2008.

Z. Lu and H. H. Ip. Constrained spectral clustering via exhaustive and efficient constraint propagation. In *European Conference on Computer Vision*, pages 1–14. Heraklion, Crete, Greece, September 2010.

Z. Lu and Y. Peng. Heterogeneous constraint propagation with constrained sparse representation. In *IEEE International Conference on Data Mining*, pages 1002–1007, Brussels, Belgium, December 2012.

Z. Lu and Y. Peng. Exhaustive and efficient constraint propagation: a graph-based learning approach and its applications. *International Journal of Computer Vision*, 103(3):306–325, July 2013a.

Z. Lu and Y. Peng. Unified constraint propagation on multi-view data. In *AAAI Conference on Artificial Intelligence*, Bellevue, Washington, United States, July 2013b.

A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *IEEE International Conference on Computer Vision*, pages 3567–3574, Sydney, Australia, December 2013.

Y. Ma, X. Hua, L. Lu, and H. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, October 2005.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

P. K. Mallapragada, R. Jin, and A. K. Jain. Active query selection for semi-supervised clustering. In *IEEE International Conference on Pattern Recognition*, pages 1–4, Tampa, Florida, United States, December 2008.

A. Margolis. A literature review of domain adaptation with unlabeled data. *Rapport Technique, University of Washington*, page 35, 2011.

J. Marin, D. Vázquez, A. M. López, J. Amores, and B. Leibe. Random forests of local experts for pedestrian detection. In *IEEE International Conference on Computer Vision*, pages 2592–2599, Sydney, Australia, December 2013.

MarketsandMarkets. Video surveillance market by system (analog, ip) by hardware, software & services, application (infrastructure, commercial, institutional, industrial and residential), and geography- global trends & forecasts to 2013 - 2020. Technical report, November 2014.

J. K. Martin. An exact probability metric for decision tree splitting and stopping. *Machine Learning*, 28(2-3):257–291, August 1997.

R. Martín-Félez and T. Xiang. Gait recognition by ranking. In *European Conference on Computer Vision*, pages 328–341. Florence, Italy, October 2012.

G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*. Marcel Dekker, 1988.

L. L. McQuitty. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educational and Psychological Measurement*, 17(2):207–229, July 1957.

M. Meila and J. Shi. A random walks view of spectral segmentation. 2001.

B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht. On oblique random forests. In *Machine Learning and Knowledge Discovery in Databases*, pages 453–469. September 2011.

A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, February 2008.

A. Montillo, J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In *Information Processing in Medical Imaging*, pages 184–196. Springer, 2011.

C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, September 2003.

B. Nelson and I. Cohen. Revisiting probabilistic models for clustering with pair-wise constraints. In *International Conference on Machine learning*, pages 673–680, Corvallis, Oregon, United States, June 2007.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, Vancouver, British Columbia, Canada, December 2002.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine learning*, pages 689–696, Bellevue, Washington, United States, July 2011.

M. S. Nixon, T. Tan, and R. Chellappa. *Human identification based on gait*. Springer, 2010.

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

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

M. W. Online-Dictionary. Cluster analysis. 2015.

O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 709–716, San Francisco, California, United States, June 2010.

S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, United States, June 2015.

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

G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, Providence, Rhode Island, United States, June 2012.

M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):167–172, January 2007.

Y. Pei, T.-K. Kim, and H. Zha. Unsupervised random forest manifold alignment for lipreading. In *IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013.

F. Perbet, B. Stenger, and A. Maki. Random forest clustering and application to video segmentation. In *British Machine Vision Conference*, London, United Kingdom, September 2009.

F. Porikli. Inter-camera color calibration by correlation model function. In *IEEE International Conference on Image Processing*, volume 2, pages II–133, Barcelona, Spain, September 2003.

V. Premachandran and R. Kakarala. Consensus of k-NNs for robust neighborhood selection on graph-based manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, United States, June 2013.

Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007.

Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1971–1984, November 2008.

Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg. Clustered synopsis of surveillance video. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 195–200, Genova, Italy, September 2009.

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

N. Quadrianto and C. H. Lampert. Learning multi-view neighborhood preserving projections. In *International Conference on Machine learning*, pages 425–432, Bellevue, Washington, United States, July 2011.

J. R. Quinlan. *C4.5: programs for machine learning*. Morgan kaufmann, 1993.

L. R. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993.

P. Rai and H. Daume. Multi-label prediction via sparse infinite cca. In *Advances in Neural Information Processing Systems*, pages 1518–1526, Vancouver, British Columbia, Canada, December 2009.

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia*, pages 251–260, Firenze, Italy, October 2010.

C. A. Ratanamahatana and E. Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*, 2004.

A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York City, New York, United States, June 2006.

M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool. Incremental learning of ncm forests for large-scale image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3654–3661, Columbus, Ohio, United States, June 2014.

M. Ristin, J. Gall, M. Guillaumin, and L. Van Gool. From categories to subcategories: Large-scale image classification with partial class label refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 231–239, Boston, Massachusetts, United States, June 2015.

S. J. Roberts, C. Holmes, and D. Denison. Minimum-entropy data clustering using reversible jump markov chain monte carlo. In *International Conference on Artificial Neural Networks*, pages 103–110. Vienna, Austria, August 2001.

M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *IEEE International Conference on Computer Vision*, pages 1235–1242, Barcelona, Spain, November 2011.

S. Rota Bulo and P. Kontschieder. Neural decision forests for semantic image labelling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–88, Columbus, Ohio, United States, June 2014.

M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE International Conference on Computer Vision*, pages 1036–1043, Barcelona, Spain, November 2011.

S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, February 2005.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, June 1990.

S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof. Alternating regression forests for object detection and pose estimation. In *IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013a.

S. Schulter, P. Wohlhart, C. Leistner, A. Saffari, P. M. Roth, and H. Bischof. Alternating decision forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, United States, June 2013b.

S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof. Accurate object detection with joint classification-regression random forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 923–930, Columbus, Ohio, United States, June 2014.

A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, Providence, Rhode Island, United States, June 2012.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138, June 2006.

Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, United States, June 2015.

R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 801–808, Colorado Springs, Colorado, United States, June 2011.

R. Sim and N. Roy. Global a-optimal robot exploration in slam. In *IEEE International Conference on Robotics and Automation*, pages 661–666, Barcelona, Spain, April 2005.

T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, December 2003.

D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *European Conference on Computer Vision Workshop*, pages 423–432, Florence, Italy, October 2012.

P. H. Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17 (1):201–226, August 1957.

A. Sobral. BGSLibrary: An opencv c++ background subtraction library. In *IX Workshop de Viso Computacional*, Botafogo, Rio de Janeiro, Brazil, June 2013.

Y. Sogawa, T. Ueno, Y. Kawahara, and T. Washio. Active learning for noisy oracle via density power divergence. *Neural Networks*, 46:133–143, October 2013.

R. R. Sokal, P. H. Sneath, et al. Principles of numerical taxonomy. *Principles of Numerical Taxonomy*, 1963.

M. G. A. Spriggs, J. Allen, M. Hemming, P. J. D. Kara, J. K. R. Little, and D. Swain. Control room operation: findings from control room observations. 2005.

N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230, Lake Tahoe, Nevada, United States, December 2012.

H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1:801–804, 1956.

A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, March 2003.

A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search*, pages 58–64, Austin, Texas, United States, August 2000.

M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3394–3401, Providence, Rhode Island, United States, June 2012.

D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *IEEE International Conference on Computer Vision*, pages 3224–3231, Sydney, Australia, December 2013.

D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3786–3793, Columbus, Ohio, United States, June 2014.

C. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. Delp. Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791, August 2006.

L. F. Teixeira and L. Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 30(2):157–167, January 2009.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv e-prints*, 2000.

G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, California, United States, June 2010.

A. Topchy, A. K. Jain, and W. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):1–37, February 2007.

R. Udupa and M. Khapra. Improving the multilingual user experience of wikipedia using cross-language name search. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 492–500, Los Angeles, California, United States, June 2010.

i-LIDS Multiple Camera Tracking Scenario Definition. UK Home Office, 2008.

T. Van Craenendonck and H. Blockeel. Limitations of using constraint set utility in semi-supervised clustering. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Porto, Portugal, September 2015.

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

A. Vinokourov, N. Cristianini, and J. S. Shawe-Taylor. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in Neural Information Processing Systems*, pages 1473–1480, Vancouver, British Columbia, Canada, December 2002.

P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, August 2007.

K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *International Conference on Machine learning*, pages 577–584, Williamstown, Massachusetts, United States, June 2001.

K. L. Wagstaff, S. Basu, and I. Davidson. When is constrained clustering beneficial, and why? In *AAAI Conference on Artificial Intelligence*, pages 62–63, Boston, Massachusetts, United States, July 2006.

H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, London, United Kingdom, September 2009.

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

J. Wang, S.-F. Chang, X. Zhou, and S. Wong. Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, United States, June 2008.

J. Wang, X. Zhu, and S. Gong. Video semantic clustering with sparse and incomplete tags. In *AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, United States, February 2016.

M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14 (4):975–985, August 2012a.

S. Wang, J. Yang, Y. Zhao, A. Cai, and S. Li. A surveillance video analysis and storage scheme for scalable synopsis browsing. In *IEEE International Conference on Computer Vision Workshop*, pages 1947–1954, November 2011.

T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*. Zurich, Switzerland, September 2014b.

X. Wang and I. Davidson. Active spectral clustering. In *IEEE International Conference on Data Mining*, pages 561–568, Sydney, Australia, December 2010.

X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, January 2012b.

X. Wang, B. Qian, and I. Davidson. Labels vs. pairwise constraints: A unified view of label propagation and constrained spectral clustering. In *IEEE International Conference on Data Mining*, pages 1146–1151, Brussels, Belgium, December 2012c.

Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtubecat: Learning to categorize wild web videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, California, United States, June 2010.

R. Waters and J. Morris. Electrical activity of muscles of the trunk during walking. *Journal of Anatomy*, 111:191–199, February 1972.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, February 2009.

K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1473–1480, Vancouver, British Columbia, Canada, December 2005.

P. Welinder and P. Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 25–32, San Francisco, California, United States, June 2010.

M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, pages 1481–1488, Vancouver, British Columbia, Canada, December 2004.

W. Wolf. Keyframe selection by motion analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, United States, May 1996.

Y. Wu, W. Li, M. Minoh, and M. Mukunoki. Can feature-based inductive transfer learning help person re-identification? In *IEEE International Conference on Image Processing*, Melbourne, Australia, September 2013.

T. Xiang and S. Gong. Spectral clustering with eigenvector selection. *Pattern Recognition*, 41 (3):1012–1029, March 2008.

J. Xiao, H. Cheng, H. Sawhney, C. Rao, and M. Isnardi. Bilateral filtering-based optical flow estimation with occlusion detection. In *European Conference on Computer Vision*, pages 211–224. Graz, Austria, May 2006.

E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, Vancouver, British Columbia, Canada, December 2002.

C. Xiong, D. Johnson, R. Xu, and J. J. Corso. Random forests for metric learning with implicit pairwise position dependence. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 958–966, San Jose, California, United States, August 2012.

Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang. *A unified framework for video summarization, browsing & retrieval: with applications to consumer and surveillance video*. Academic Press, 2006.

Q. Xu, K. L. Wagstaff, et al. Active constrained clustering by examining spectral eigenvectors. In *Discovery Science*, volume 3735, pages 294–307, October 2005.

Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013.

O. Yakhnenko and V. Honavar. Multiple label prediction for image annotation with multiple kernel correlation models. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 8–15, Miami, Florida, United States, June 2009.

Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from crowds. In *International Conference on Machine learning*, pages 1161–1168, Bellevue, Washington, United States, July 2011.

H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *IEEE International Conference on Computer Vision*, pages 1936–1943, Sydney, Australia, December 2013.

L. Yang and R. Jin. Distance metric learning: A comprehensive survey. Technical report, Michigan State Universiy, May 2006.

Z. Yang, Y. Hu, H. Liu, H. Chen, and Z. Wu. Matrix completion for cross-view pairwise constraint propagation. In *ACM International Conference on Multimedia*, pages 897–900, Orlando, Florida, United States, November 2014.

B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1577–1584, Colorado Springs, Colorado, United States, June 2011.

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

P. Yin, A. Criminisi, J. Winn, and I. Essa. Tree-based classifiers for bilayer video segmentation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8, Minneapolis, Minnesota, United States, June 2007.

Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research*, 13(1):1–26, January 2012.

G. Yu, J. Yuan, and Z. Liu. Unsupervised random forest indexing for fast action search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 865–872, Colorado Springs, Colorado, United States, June 2011.

S. X. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, February 2004.

L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1601–1608, Vancouver, British Columbia, Canada, December 2004.

E. Zeng, C. Yang, T. Li, and G. Narasimhan. On the effectiveness of constraints sets in clustering genes. In *IEEE International Conference on Bioinformatics and Bioengineering*, pages 79–86, Boston, Massachustts, United States, October 2007.

D.-Q. Zhang, C.-Y. Lin, S.-F. Chang, and J. R. Smith. Semantic video clustering across sources using bipartite spectral clustering. In *IEEE International Conference on Multimedia and Expo*, Sorrento, Italy, June 2004.

H. Zhang, J. Wu, D. Zhong, and S. W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, April 1997.

X. Zhang, L. Zong, X. Liu, and H. Yu. Constrained nmf-based multi-view clustering on unmapped data. In *AAAI Conference on Artificial Intelligence*, Austin, Texas, United States, January 2015.

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

R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, Portland, Oregon, United States, June 2013b.

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

R. Zhao, W. Ouyang, and X. Wang. Person re-identification by saliency learning. In *arXiv e-prints*, 2014b.

X. Zhao, T.-K. Kim, and W. Luo. Unified face analysis by iterative multi-output random forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1772, Columbus, Ohio, United States, June 2014c.

Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, June 2004.

L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image

search and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, United States, June 2015.

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

W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *British Machine Vision Conference*, London, United Kingdom, September 2009.

D. Zhou and C. J. Burges. Spectral clustering and transductive learning with multiple views. In *International Conference on Machine learning*, pages 1159–1166, Corvallis, Oregon, United States, June 2007.

D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *Advances in Neural Information Processing Systems*, pages 169–176, Vancouver and Whistler, British Columbia, Canada, December 2004.

X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, December 2005.

X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine learning*, volume 3, pages 912–919, Washington D.C.? United States, August 2003.

X. Zhu, C. C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation. In *IEEE International Conference on Computer Vision*, pages 81–88, Sydney, Australia, December 2013.

X. Zhu, C. C. Loy, and S. Gong. Constructing robust affinity graphs for spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1450–1457, Columbus, Ohio, United States, June 2014.