METHODS FOR LARGE-SCALE MACHINE LEARNING AND COMPUTER VISION

by

YEQING LI

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2015

To my family, for their endless trust, support, encourage and love.

ACKNOWLEDGEMENTS

There were many people who helped me during my PhD studying career, and I would like to take this opportunity to thank them.

I would like to thank my supervising professor Dr. Junzhou Huang for constantly motivating and encouraging me, and also for his invaluable advice during the course of my doctoral studies. He held me to the highest of standards, but also had the faith that I would be able to achieve them. None of the work in this thesis would have happened without him.

I wish to thank my thesis committee members Dr. Heng Huang, Dr. Chris Ding, Dr. Jeff Lei and Dr. Dimitris Metaxas for their interest in my research and for their valuable suggestions regarding my early proposal and this thesis. It is a privilege for me to have each of them serve in my committees

The research described in this thesis has benefited from my other collaborators besides my advisors. Without them, some of the chapters in this thesis would not have been possible. My special thanks go to Dr. Feiping Nie, Dr. Wei Liu, Prof. Leon Axel and Prof. Xiaolei Huang. I have been learning a lot from them through the collaborations.

I want to thank all my colleagues from the Scalable Modeling and Imaging and Learning Lab (SMILE), the Computer Science and Engineering Department. It is my pleasure to meet such a concentration of creative and nice people here. I am grateful to all with whom I spent my time as a graduate student at UTA.

Finally, my special thanks go to my family. I would like to express my earnest gratitude to my parents for their love and countless sacrifices to give me the best pos-

sible education. I cannot thank my parents-in-law enough for giving me the opportunity to freely pursue my interests and career. Without their patience and unreserved support, it would not have been possible to reach this stage in my career. I owe a debt of gratitude to my wife Shaojie Cai who have loved and supported me throughout this thesis and throughout my life. This thesis is built on the sacrifice of her career in China. She made my life is never entirely consumed by research. Finally, thanks to my little boy Alan, for the energy and the happiness he bring to me.

September 1, 2015

ABSTRACT

METHODS FOR LARGE-SCALE MACHINE LEARNING AND COMPUTER VISION

YEQING LI, Ph.D.

The University of Texas at Arlington, 2015

Supervising Professor: Junzhou Huang

With the advance of the Internet and information technology, nowadays people can easily collect and store tremendous amounts of data such as images and videos. Developing machine learning and computer vision to analysis and learn from the gigantic data sets is an interesting yet challenging problem. Inspired by the trend, this thesis focus on developing large-scale machine learning and computer vision techniques for the purpose of handling various kinds of problems on gigantic data sets.

With respect to the problem of image classification, we employ the technique of sub-selection, which uses partial observations to efficiently approximate the original high dimensional problems.. We consider the classification models based on sparse representation or collaborative representation. In practical applications, the performance of classification can be affected by problems like misalignment, occlusion and big noises. To deal with these problems, we propose a robust sub-representation method, which can effectively handle these problems with an efficient scheme.

With respect to the problem of similarity search, this thesis contribute a novel method for hashing a large number of images. While many researchers have worked on

the topic of how to find good hash function for this task, the thesis will propose a new approach to address efficiency. In particular, the training step of many existing hash methods relies on computing the Principle Components Analysis (PCA). However, performing PCA on large dataset is time-consuming. The thesis will prove that, under some conditions, the PCA can be computed by using only a small part of the data. With the theoretical guarantee, one can accelerate the training process of hashing without loss much of accuracy.

With respect to the problem of large-scale multi-view clustering, the thesis contribute a novel method for graph-based clustering. A graph offers an attractive way of representing data and discovering the essential information such as the neighborhood structure. However, both of the graph construction process and graph-based learning techniques become computationally prohibitive at a large scale. To overcome these bottlenecks, we present a novel graph construction approach, called Salient Graphs, which enjoys linear space and time complexities and can thus be constructed over gigantic databases efficiently. Then, we implement an efficient graph-cut algorithm, which iteratively search consensus between multiple views and perform clustering. This results in an accurate and fast algorithm for multi-view data clustering.

With respect to the problem of visual tracking, the thesis contribute a novel method for instrument tracking in retinal microsurgery. The instrument tracking is a key task in robot-assist surgical system. In this kind of system, data is collected and processing in real-time. Therefore, a tracking algorithm need to find good balance between accuracy and efficiency. The thesis proposed a novel visual tracker based on online learning. The proposed algorithm is able to run in video frame-rate while achieving the state-of-the-art accuracy.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This thesis focus on developing large-scale machine learning and computer vision techniques for the purpose of handling supervised, semi-supervised and unsupervised problems, e.g. classification, clustering, nearest neighbor search.

1.1 Motivation

In the current era of big data, more and more efforts has been devoted to leveraging massive amounts of data available in open sources such as Internet to help solve various computer vision, data mining and information retrieval problems. Within this context, how to effectively extract knowledge from and efficiently exploit large-scale data is still an open problem. In this thesis, we aim at providing effective machine learning and computer vision approaches for handling large-scale data sets.

First of all, we explore the sparse representation [1] and collaborative representation [2] for the classification of high dimensional image data. The sparse representation offers an attractive way to represent image data and perform classification. The basic idea of sparse representation is that each data point can be represented by the linear combination of the samples in the dictionary with one additional constrain that the representation coefficients must be sparse (usually by using ℓ_1-norm regularization). Also, several extensions have been made to sparse representation model to handle problems like image transformation, occlusions. These method have demonstrated promising results on various tasks such as visual tracking, face recognition. However, one drawback of sparse representation is that it require the dictionary to

be over complete. In other word, one must collect enough training data. Zhang et al. proposed collaborative representation [2], [3] which relax the sparsity constraint on the representation coefficients. CR inspires many following works, such as MRR [4], DLRD-SR [5]. Collaborative representation uses the ℓ_2-norm regularization instead of ℓ_1-norm in sparse representation. With the relaxation, collaborative representation can achieve better performance than the sparse representation when the dictionary is under-complete. However, both the sparse and collaborative approaches are ineffective in handling high dimensional data.

Secondly, we investigate the problem of handling large-scale high-dimensional data. Specifically, we focus on approximate nearest neighbor (NN) search for large-scale high-dimensional data. We mainly focus on the emerging technique called hashing, which is becoming increasingly popular for time-efficient NN search on multimedia data, especially image and video. It has been shown that mapping high-dimensional image descriptors to compact binary codes can lead to considerable efficiency gains in both storage and similarity computation of images. However, most existing methods still suffer from expensive training devoted to large-scale binary code learning. Therefore, we are trying to develop an approach to reduce the computational cost the training process.

All of the above problems are supervised learning problems. Besides that, we are also interested in developing efficient algorithm for unsupervised learning problems. Among them, clustering is the one of the fundamental one. We also try to address the clustering problem on large-scale multi-view data. This is motivated by the following fact: in many real-world applications, data can be represented in various heterogeneous features or views. Different views often provide different aspects of information that are complementary to each other. Several previous methods of clustering have demonstrated that better accuracy can be achieved using integrated information of

all the views than just using each view individually. One important class of such methods is multi-view spectral clustering, which is based on graph Laplacian. On drawback of the existing methods is their high computational complexity.

The main focus of the previous problems is how to increase the throughput of the algorithm. Other than that, there are also broad interest in developing algorithm to handle stream of data in real-time. Visual tracking is one of this kind of problems, which try to locate the given target object in a video sequence. Here, we study the problem of instrument tracking in retinal microsurgery. Robust visual tracking of instruments is an important task in retinal microsurgery. In this context, the instruments are subject to a large variety of appearance changes due to illumination and other changes during a procedure, which makes the task very challenging. Most existing methods require collecting a sufficient amount of labelled data and yet perform poorly in handling appearance changes that are unseen in training data. How to effectively process the video sequence in real-time is a challenging problem.

1.2 Our Techniques

There is no single simple answer to the question of how to achieve accuracy and efficiency in machine learning tasks on multimedia databases. The answer depends on many factors, including the specific types of data that we are dealing with (e.g., video, audio, biological sequences), the problems we are trying to solve (unsupervised, semi-supervised or supervised) and the environment where the algorithm is deployed (e.g., off-line batch learning or real-time on-line learning). A large body of literature exists that describes various methods for different variations of the problem.

First, we present the idea of sub-representation that is used to accelerate the sparse representation and the collaborative representation. The proposed sub-representation method that can handle misalignment, occlusion and big noises with

lower computational cost. It is motivated by the sub-selection technique, which uses partial observations to efficiently approximate the original high dimensional problems. In practical applications, classification performance is affected by problems like misalignment, occlusion and big noises. Therefore, we propose a robust sub-representation method, which can effectively handle these problems with an efficient scheme. While its performance guarantee was theoretically proved, numerous experiments on practical applications have further demonstrated that the proposed method can lead to significant performance improvement in terms of speed and accuracy.

Secondly, for the image searching topic, we propose a sub-selection based matrix manipulation algorithm which can significantly reduce the computational cost of code learning. This is based on the observation that many existing methods relies on projecting data to low dimensional space, which usually involves slow matrix operations. We demonstrate that the most time-consuming matrix operations encountered in code learning, typically data projection and rotation, can be performed in a more efficient manner. A fast matrix multiplication algorithm is proposed using a sub-selection [6] technique to accelerate the learning of coding functions. Our algorithm is motivated by the observation that the degree of the algorithm parameters is usually very small compared to the number of entire data samples. Therefore, we are able to determine these parameters merely using partial data samples As case studies, we apply the sub-selection algorithm to two popular quantization techniques PCA Quantization (PCAQ) and Iterative Quantization (ITQ). Crucially, we can justify the resulting sub-selective quantization by proving its theoretic properties.

Thirdly, for the clustering topic, we focus on the spectral clustering. The spectral clustering is based on the neighborhood graph of the data. A graph offers an attractive way of representing data and discovering the essential information such as the neighborhood structure. However, both of the graph construction process and

graph-based learning techniques become computationally prohibitive at a large scale. We present an efficient spectral clustering algorithm for large-scale multi-view data by using a Conventional neighborhood graphs such as kNN graphs require a quadratic time complexity, which is inadequate for large-scale applications mentioned above. To overcome this bottleneck, we present a novel graph construction approach, called Salient Graphs, which enjoys linear space and time complexities and can thus be constructed over gigantic databases efficiently. The central idea of the Salient Graph is introducing a few salient points and converting intensive data-to-data affinity computation to drastically reduced data-to-salient affinity computation. A low-rank data-to-data affinity matrix is derived using the data-to-salient affinity matrix. Then we construct bipartite graph between raw data points and these salient points. Then, the graph of all the views are combined together using a local manifold fusion method. Finally, we run spectral clustering on the resulting fused graph. There are several benefits of our method: **First**, manifold fusion preserves the manifold structure of all the views; **Second**, the construction of the bipartite graph is very efficient; **Third**, by exploring the special structure of the bipartite graph, spectral analysis on it is also very efficient; **Fourth**, our method also output cluster indicator of the salient points, which enables us to handle the out-of-sample problem efficiently. Additionally, we have conducted extensive experiments on five benchmark data sets, which demonstrate the effectiveness and efficiency of our proposed method comparing to the state-of-the-art methods.

Finally, we proposed an online learning approach for visual tracking task in computer-assisted surgical. This is a different scenario which requires data to be handled in real-time. Robust visual tracking of instruments is an important task in retinal microsurgery. In this context, the instruments are subject to a large variety of appearance changes due to illumination and other changes during a procedure,

which makes the task very challenging. Most existing methods require collecting a sufficient amount of labeled data and yet perform poorly in handling appearance changes that are unseen in training data. To address these problems, we propose a new approach for robust instrument tracking. In this approach, we adopt the paradigm of combining tracking and detection in the same framework. The proposed approach uses a robust gradient-based tracker capable of failure detection as the basic tracker. Then, a cascade appearance classifier is used as the instrument detector. The appearance model of the detector is initialized by manually clicking the instrument position in the first frame. It is adaptively trained and updated on the fly. Samples for online updating are collected by a filtering process, which selects "unfamiliar" positive samples and "hard" negative samples. The obtained training set is used to augment the model of the detector and prevent the detector from making the similar mistakes.

1.3 Thesis Overview

Finally, we provide the overview of this thesis in brief. In Chapter 2, we present our sub-representation approach to handle high-dimensional data. Then, Chapter 3 generalize the sub-representation to image hashing problem to handle large-scale high-dimensional data. Chapter 4 turns to unsupervised learning problem on another kind of data: large-scale multi-view data. Then, Chapter 5 presents the online algorithm for the instrument tracking problem, which is a typical example of problem on handling stream data.

As the ending, Chapter 6 draws our conclusions of the thesis, where we summarize the presented large-scale machine learning and computer vision techniques, highlight their contributions in both theory and practice, and provide some future research directions.

CHAPTER 2

TRANSFORMATION-INVARIANT COLLABORATIVE SUB-REPRESENTATION

This chapter investigates the problem of handling high-dimensional data in image representation problem. A novel approach named sub-representation is proposed. Theoretical analysis is also provided for the performance of the proposed approach [6].

2.1 Introduction

Image representation is an important problem in computer vision and pattern recognition and it has gained a lot of attentions in past decade. While huge interest has been seen in image representation, to date the most popular approaches are sparse representation and collaborative representation.

In Wright et al.s pioneer work SRC [1] on sparse representation, they model the recognition problem as finding sparse representation of the test image based on linear combination of training images. Furthermore, outlier pixels are also assumed to be sparse. Favourable result has been achieved on face recognition application with occlusion and corruption [7, 8, 9]. Zhang et al. proposed collaborative representation (CR) [2], [3] which relax the sparsity constraint on the representation coefficients. CR inspires many following works, such as MRR [4], DLRD SR [5]. CR uses the ℓ_2-norm regularization instead of ℓ_1-norm in SRC. Therefore, it is much faster. It can achieve desired results for the clean data without corruption. However, if the testing images are sparsely corrupted, ℓ_1-norm regularization has to be used for constraining

sparse occlusion, which will lead to impressively degradation of speed. The high computation complexity is a big obstacle for them being used for high-dimensional images.

Moreover, in practice, due to the registration error of the object detector or the motion of the target object, the test image is usually not well-aligned with the training images. To achieve effective representation, the test image has to be aligned with training images first by using iterative transformation estimation methods [4], [10], [11]. This process has higher computation costs because the transformation estimation step is usually solved in high dimensional pixel-space and not able to utilize dimension reduction techniques. Though existing methods have achieved great success in various situations such as illumination change, occlusion and misalignment, their computational inefficiency limits them being used in practical application involving high-dimensional images.

To this end, we propose a robust sub-representation method, which can not only efficiently represent the image but also effectively handle the problems of misalignment, occlusion and big noises. Its performance guarantee can be theoretically proved. While combining it with existing collaborate representation method, a Transformation-invariant Collaborative Sub-Representation (TCSR) algorithm is proposed in this chapter. Numerous experiments on practical applications have been conducted to further demonstrate its superior performance in terms of both computational complexity and accuracy.

The contributions of this chapter are: **1)** To handle big noises, sub-selection method is generalized to robust sub- representation. We have theoretically proved its benefit over sub-selection method; **2)** Sub-representation is further extended to handle misalignment, occlusion and corrupted pixels, etc.. This extension is done by combing it with some existing techniques like transformation estimation algorithm

and collaborative representation. The resulting method can also be regarded as a case study of sub-representation, which shows its great potential in cooperating with other methods. **3)** Extensive experiments are conducted to validate the efficiency and effectiveness of the proposed methods, which demonstrates that the proposed method can significantly accelerate the collaborative representation with imperceptible loss of accuracy.

2.2 Related Work

2.2.1 Sub-selection and Handling Incomplete Data

Sub-selection [12] is an efficient way to reduce the dimension of data. It projects the data onto lower dimensional subspaces using a randomly chosen subset of the data features. An example of sub-selection on pixel features of an image is shown in Fig. 2.1. In theoretical analysis, one situation that is similar to sub-selection is handling incomplete data. Balzano et. al. have recently proved theoretical guarantees of subspace detection using incomplete data [13].

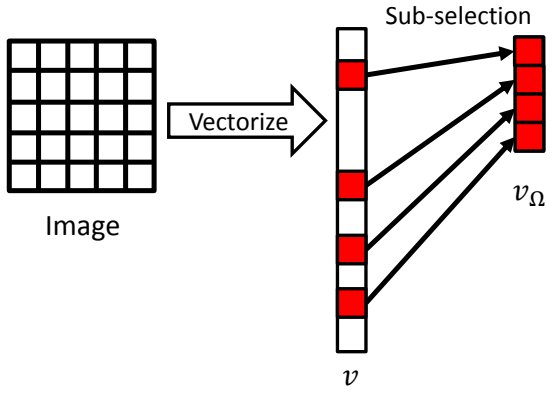


Figure 2.1: Sub-selection on pixel features of an image.

The theory is sketched as follows for convenience. Let v_Ω be the vector of dimension $|\Omega| \times 1$ comprised of the elements $v_i, i \in \Omega$, ordered lexigraphically, where $|\Omega|$ denotes the cardinality of Ω. The energy of v in the subspace S is $\|P_S v\|_2^2$, where P_S denotes the projection operator onto S. Let U be an $n \times r$ matrix whose columns span the r-dimensional subspace S. In this case, $P_S = U(U^T U)^{-1} U^T$. With this representation in mind, let U_Ω denote the $|\Omega| \times r$ matrix, whose rows are the $|\Omega|$ rows of U indexed by the set Ω, arranged in lexigraphic order. Suppose we only observe v on the set Ω. One approach for estimating its energy in S is to assess how well v_Ω can be represented in terms of the rows of U_Ω. Define the projection operator $P_{S_\Omega} := U_\Omega (U_\Omega^T U_\Omega)^\dagger U_\Omega^T$, where $\dagger$ denotes the pseudo-inverse. It follows immediately that if $v \in S$, then $\|v - P_S v\|_2^2 = 0$ and $\|v_\Omega - P_{S_\Omega} v_\Omega\|_2^2 = 0$.

Let the entries of v be sampled uniformly with replacement. Again let Ω refer to the set of indices for observations of entries in v, and denote $|\Omega| = k$. The following theorem has been proved in [13]:

Theorem 1. *(Therom 1 in [13])*

Let $\delta > 0$ and $k \geq \frac{8}{3} r \mu(S) \log(\frac{2r}{\delta})$. Then with probability at least $1 - 4\delta$,

$$\frac{k}{n}(1 - \alpha - \alpha_0)\|v - P_S v\|_2^2 \leq \|v_\Omega - P_{S_\Omega} v_\Omega\|_2^2 \leq \frac{k}{n}(1 + \alpha)\|v - P_S v\|_2^2, \tag{2.1}$$

where $\alpha = \sqrt{\frac{2\mu(y)^2}{k} \log(\frac{1}{\delta})}$, $\beta = \sqrt{2\mu(y) \log(\frac{1}{\delta})}$, $\gamma = \sqrt{\frac{8r\mu(S)}{3k} \log(\frac{1}{\delta})}$, and $\alpha_0 = \frac{r\mu(S)}{k} \frac{(1+\beta)^2}{1-r}$.

$\mu(S)$ is the coherence of a subspace S [14]: $\mu(S) := \frac{r}{n} \max_j \|P_S e_j\|_2^2$, where e_j represents a standard basis element. Note that $1 \leq \mu(S) \leq \frac{n}{r}$. We let $\mu(y)$ denote the coherence of the subspace spanned by y. By plugging in the definition, we have $\mu(y) = \frac{n\|y\|_{\inf}^2}{\|y\|_2^2}$. Here, $v = x + y, x \in S$ and $y \in S^\perp$.

For general cases, Balzano et. al. have shown that if $|\Omega|$ is just slightly greater than $r \log(r)$, then with high probability $\|v_\Omega - P_{S_\Omega} v_\Omega\|_2^2$ is very close to $\frac{|\Omega|}{n}\|v - P_{S_v}\|_2^2$.

It means that partial observations can efficiently approximate the full observation with high dimension. Balzano et. al.s result provides a useful starting point to analyse performance of sub-selection representation. However, it only concerns clean data and can not directly extend to handle many real problems in practical applications, such as the earlier noted misalignment, occlusion and big noises.

2.2.2 Transformation-invariant Collaborative Representation

Yang et al [4] proposed a transformation-invariant collaborative representation that can handle the representation and the transformation estimation simultaneously. First, a sparse term e is employed to handle occlusion. The problem is solved by minimizing the objective function:

$$\min_{x,e} F_1(x,e) = \|y - Ax - e\|_2^2 + \lambda\|x\|_2^2 + \gamma\|e\|_1, \tag{2.2}$$

where e represents the sparse big error, $y \in \mathbb{R}^2$ is the query image, $A = [I_1, I_2, ..., I_p] \in \mathbb{R}^{n\times p}$ is the dictionary, $x \in \mathbb{R}^p$ denotes the representation coefficients. And then the transformation parameter is introduced so that the objective function becomes:

$$\min_{x,e,\tau} F_2(x,e) = \|y \odot \tau - Ax - e\|_2^2 + \lambda\|x\|_2^2 + \gamma\|e\|_1, \tag{2.3}$$

where τ is the transformation parameters and $y \odot \tau$ means to apply the transformation on the query image. The authors use a two-step strategy to accelerate the optimization process. Step one is to calculate SVD decomposition on the dictionary and use the singular vectors to solve x and τ :

$$\min_{\beta,e,\tau} F_3(\beta,e,\tau) = \|y \odot \tau - U\beta - e\|_2^2 + \gamma\|e\|_1, \tag{2.4}$$

where $A = USV^T$ is the SVD decomposition of A and $\beta = SV^Tx$ is temporal representation coefficient with big absolute values. Minimizing this $F_3(\beta, e, \tau)$ will give an

estimation of transformation parameters τ, big sparse error e, and the representation coefficients.

This approach is much faster than the previous sparse representation approach RASR [11]. However, due to the ℓ_1-norm optimization and transformation estimation, it is still slower for high-dimensional images in practical applications.

2.3 Methodology

In this section, we introduce sub-representation approach and provide the theoretical proof of its performance guarantee for the case with sparse big noise. The resulting method is called as robust sub-representation. Finally, after combining it with transformation estimation and collaborative representation, we propose the TCSR approach.

In the following discussion, we shall interchangeably use τ to indicate the transformation parameters and an operator to apply that transformation on a set of images. Likewise, we may also use Ω as a sub-selection matrix as well as a sub-selection operation using that matrix. So that $y \odot \tau$ indicates applying the transformation on image y, and $A \odot \tau$ indicates applying the transformation on each image (column) in A. Similarly, $y \odot \Omega$ and $A \odot \Omega$ represents applying sub-selection on y and A respectively.

2.3.1 Sub-Representation

Consider the problem of representing a query image by linear combination of a set of images. Let $y \in \mathbb{R}^n$ be the query image, $A = [I_1, I_2, ..., I_p] \in \mathbb{R}^{n \times p}$ be the dictionary of p images, $x \in \mathbb{R}^p$ denotes the representation coefficients. This problem can be formulated as solving the equation $Ax = y$. The dimension of this equation

can be reduced using sub- selection that we discussed in Section II-A. Instead of solving $Ax = y$, we solve $A_\Omega \hat{x} = y_\Omega$. That is

$$\hat{x} = \arg\min_x \|y_\Omega - A_\Omega x\|_2^2, \tag{2.5}$$

where $A_\Omega = A \odot \Omega$ denotes the dictionary under sub- selection, $y_\Omega = y_\Omega$ denotes the query image under sub- selection and $\hat{x}$ is the representation coefficient vector. From Theorem 1, $\hat{x}$ should be very close to original x when Ω satisfy certain conditions. Equation (2.5) is our basic idea of sub- representation. The benefit of sub-representation is the solution of the original equation can be approximated by the solution of the low dimensional version of the equation. Hence, the the computational cost is significantly reduced. From now on, we shall extend sub-representation to handle several challenging problems and finally reach a practical image representation method.

2.3.2 Robust Sub-Representation

Occlusions and corrupted pixels are common challenges in many practical scenarios. In previous literatures, they are modelled as sparse big noise. Instead of solving $Ax = y$, we need to solve $Ax = y - e$, where e is the sparse error term. Adding sparse regularization on e, the problem becomes minimizing the following objective function:

$$\min_{x,e} F_4(x, e) = \gamma\|e\|_1 + \|y - Ax - e\|_2^2, \tag{2.6}$$

where γ is the regularization parameter. Solving this equation directly can be time consuming for even a medium size image, like 256×256, due to the ℓ_1-norm minimization. However, usually $p \ll n$ or the rank of matrix A is far less than p and n, a

sub-selection operation (e.g. $|\Omega| = m \ll n$) Ω can be applied on the representation. The new objective function is:

$$\min_{\hat{x}, e_\Omega} G_4(\hat{x}, e_\Omega) = \gamma \|e_\Omega\|_1 + \|y_\Omega - A_\Omega \hat{x} - e_\Omega\|_2^2, \tag{2.7}$$

where $e_\Omega = e \odot \Omega$ is the error term resulting from acting sub-selection on e.

Now we shall discuss the relationship between Eq. (2.6) and Eq. (2.7), which is missing in the previous literatures. We first prove the boundedness of ℓ_1-norm term under sub-selection. The follow theorem is required in later discussion:

Theorem 2. *(McDiarmids Inequality [15]): Let $X_1, ..., X_n$ be independent random variables, and assume f is a function for which there exist t_i, $i = 1, \ldots, n$ satisfying*

$$\sup_{x_1, \ldots, x_n, \hat{x}+i} |f(x_1, \ldots, x_n) - f(x_1, \ldots, \hat{x}_i, \ldots, x_n)| \leq t_i \tag{2.8}$$

where $\hat{x}_i$ indicates replacing the sample value x_i with any other of its possible values. Call $f(x_1, \ldots, x_n) := Y$. Then for any $\epsilon > 0$,

$$\mathbb{P}[Y \geq \mathbb{E}[Y] - \epsilon] \leq \exp(\frac{-2\epsilon}{\sum_{i=1}^n t_i^2}) \tag{2.9}$$

$$\mathbb{P}[Y \leq \mathbb{E}[Y] + \epsilon] \leq \exp(\frac{-2\epsilon}{\sum_{i=1}^n t_i^2}) \tag{2.10}$$

With Theorem 2, the following lemma can be proved:

Lemma 1. *Suppose $\delta > 0$, $y \in \mathbb{R}^n$ and $|\Omega| = m$, then*

$$(1 - \alpha_1)\frac{m}{n}\|y\|_1 \leq \|y_\Omega\|_1 \leq (1 - \alpha_1)\frac{m}{n}\|y\|_1 \tag{2.11}$$

with probability at least $1 - 2\delta$.

Proof. We use McDiarmids inequality from Theorem 2 for the function $f(X_1, ..., X_m) = \sum_{i=1}^m X_i$ to prove this. Set $X_i = |y_\Omega(i)|$. Since $\|y_\Omega(i)\| \leq \|y\|_{\inf}$ for all i, we have $|\sum_{i=1}^m X_i - \sum_{i \neq k} X_i - \hat{X}_k| = |X_k - \hat{X}_k| \leq 2\|y\|_{\inf}$. We first calculate $\mathbb{E}[\sum_{i=1}^m X_i]$ as

follows. Define **1** to be the indicator function, and assume that the samples are taken uniformly with replacement.

$$\begin{aligned}\mathbb{E}[\sum_{i=1}^{m} X_i] &= \mathbb{E}[\sum_{i=1}^{m} |y_{\Omega(i)}|] \\ &= \sum_{i=1}^{m}[\mathbb{E}[\sum_{j=1}^{n} |y_j|\mathbf{1}_{\Omega(i)=j}]] \\ &= \frac{m}{n}\|y\|_1.\end{aligned}$$

Invoking the Theorem 2, the left hand side is

$$\begin{aligned}&\mathbb{P}[\sum_{i=1}^{m} X_i \leq \mathbb{E}[\sum_{i=1}^{m} X_i] - \epsilon] \\ =&\mathbb{P}[\sum_{i=1}^{m} X_i \leq \frac{m}{n}\|y\|_1 - \epsilon].\end{aligned}$$

We can let $\epsilon = \alpha\frac{m}{n}\|y\|_1$ and then have that this probability is bounded by

$$\exp\left(\frac{-2\alpha^2(m)^2\|y\|_1^2}{4m\|y\|_{\text{inf}}^2}\right).$$

Thus, the resulting probability bound is

$$\mathbb{P}[\|y_\Omega\|_1 \geq (1-\alpha)\frac{m}{n}\|y\|_1] \geq 1 - \exp\left(\frac{-\alpha^2 m\|y\|_1^2}{2n^2\|y\|_{\text{inf}}^2}\right).$$

Substituting our definitions of $\mu_1(y) = \frac{n\|y\|_{\text{inf}}}{\|y\|_1}$ and $\alpha_1 = \sqrt{\frac{2\mu_1(y)^2}{m}\log(\frac{1}{\delta})}$ shows that the lower bound holds with probability at least $1-\delta$. The argument for the upper bound can be proved similarly. The Lemma now follows by applying the union bound. □

Lemma 2. *Let $\delta > 0$. Then with probability at least $1-6\delta$, $F_4(x,e)$ in Eq. (2.6) and $G_4(\hat{x}, e_\Omega)$ in Eq. (2.7) satisfy:*

$$\frac{m}{n}(1-\alpha_4)F_4(x,e) \leq G_4(\hat{x}, e_\Omega) \leq \frac{m}{n}(1+\alpha_4)F_4(x,e), \tag{2.12}$$

where a_4 is a small positive constant for given problem.

Proof. Lemma 2 can be proved using Theorem 1 and Lemma 1. F_4 and G_4 are objective functions with two unknowns. Let t_1 the first term of F_3 and t_2 be the first term of G_4, we have

$$t_1 = \|\tilde{y} - Ax\|_2^2, \tag{2.13}$$

where $\tilde{y} = y - e$, $C_1 = \lambda\|e\|_1$, and

$$t_2 = \|\tilde{y}_\Omega - A_\Omega \hat{x}\|_2^2, \tag{2.14}$$

where $\tilde{y}_\Omega = y_\Omega - e_\Omega$, $C_2 = \lambda\|e_\Omega\|_1$. With Theorem 1, we have:

$$\frac{m}{n}(1 - \alpha - \alpha_0)T_1 \leq T_2(\hat{x}) \leq \frac{m}{n}(1 + \alpha)T_1 s(x), \tag{2.15}$$

with probability at least $1 - 4\delta$.

Similarly, we can prove the second term of G_4 and F_4 are bounded with probablity $1 - 2\delta$ using Lemma 1. Hence, the bound of G_4 and F_4 can be proved by applying the union bound of these two parts. □

With Lemma 2, its easy to derive that the sub- representation coefficient $\hat{x}$ will be very close to the origin solution x. Lemma 2 provides the theoretical guarantee for quality of solution of Eq. (2.7).

2.3.3 Sub-selection and Transformation Estimation

Sub-selection can be used to accelerate transformation estimation. It can be formulated as:

$$\tau = \arg\min_{\tau} \|y_1 - y_2 \odot \tau\|_2^2, \tag{2.16}$$

where y_1, y_2 are two images without correct alignment, τ is the unknown parameter of the transformation. For many kinds of transformation, like affine, similarity, homograph, translation, etc., the formula can be solved by an iterative approach [16].

To approximate Eq. (2.16), a Taylor expansion is usually applied. The problem becomes minimizing:

$$\min_{\Delta\tau} F_5(\Delta\tau) = \|y_1 - y_2 \odot \tau_c - J_\tau \Delta\tau\|_2^2, \tag{2.17}$$

where τ_c is the current estimation of the transformation, $\Delta\tau$ is the parameter increment at each iteration and J_τ is the Jacobian matrix. Eq. (2.17) is a least square equation and have close form solution. The dimension of the transformation parameters is usually very low compared to the image dimension. Hence, sub-selection is applicable here. Assume Ω is a sub-selection matrix(n choose m), $y_{\Omega,1} = y_1 \odot \Omega, y_{\Omega,2} = y_2 \odot \Omega$ and $J\tau, \Omega = Jacobian_{\hat{\tau}(y\Omega,2)}$. The result objective function is as follow:

$$\min_{\Delta\hat{\tau}} G_5(\Delta\hat{\tau}) = \|y_{\Omega,1} - y_{\Omega,2} \odot \hat{\tau}_c - J_{\hat{\tau},\Omega} \Delta\hat{\tau}\| \tag{2.18}$$

Then, we have:

Lemma 3. *Let $\delta > 0$. Then with probability at least $1 - 4\delta$, $F_5(\Delta\tau)$ ing Eq. (2.17) and $G_5(\Delta\hat{\tau})$ in Eq. (2.18) satisfy:*

$$\frac{m}{n}(1 - \alpha_5)F_5(x, e) \leq G_5(\hat{x}, e_\Omega) \leq \frac{m}{n}(1 + \alpha_4)F_4(x, e),$$

where a_5 is a small positive constant for given problem.

Proof. Lemma 3 can be proved using Theorem 1. We first transform F_5 and G_5 as follows:

$$F_5 = \|y_1 - y_2 \odot \tau_c - J_\tau \Delta\tau\|_2^2 = \|t_1 - J_\tau \Delta\tau\|_2^2, \tag{2.19}$$

$$G_5 = \|y_{\Omega,1} - y_{\Omega,2} \odot \tilde{\tau}_c - J_{\tilde{\tau}} \Delta\tilde{\tau}\|_2^2 = \|t_2 - J_{\tilde{\tau}} \Delta\tilde{\tau}\|_2^2, \tag{2.20}$$

where $t_1 = y_1 - y_2 \odot \tau_c$ and $t_2 = y_{\Omega,1} - y_{\Omega,2} \odot \tilde{\tau}_c = t_{1,\Omega}$. The Lemma is proved by applying Theorem 1. □

With this Lemma, $\Delta\hat{\tau}$ should be very close to $\Delta\tau$. This property allows us to address the unresolved problem of solving transformation estimation in the same low dimensional space of solving the representation coefficients. Lemma 3 provides theoretical guarantee for quality of solution of Eq. (2.18).

2.3.4 Transformation Invariant Collaborative Sub-Representations (TCSR)

Now, we complete our discussion by combining techniques we discuss above and reach our final proposed method. The basic idea is transform Eq. (3) to low dimensional space via sub-selection. There resulting formula is as follow:

$$\min_{\hat{x},e_\Omega,\hat{\tau}} G_2(\hat{x},\hat{e},\hat{\tau}) = \|y_\Omega \odot \tau - A_\Omega x - e_\Omega\|_2^2 + \lambda\frac{m}{n}\|\hat{x}\|_2^2 + \gamma\|e_\Omega\|_1, \tag{2.21}$$

where $\hat{x}$, $\hat{\tau}$ are the representation parameters and transforma- tion parameters respectively. Then we have:

Lemma 4. *Let $\delta > 0$. Then with probability at least $1 - 6\delta$, $F_2(\Delta\tau)$ ing Eq. (2.3) and $G_2(\Delta\hat{\tau})$ in Eq. (2.21) satisfy:*

$$\frac{m}{n}(1-\alpha_7)F_2(x,e) \leq G_2(\hat{x},e_\Omega) \leq \frac{m}{n}(1+\alpha_7)F_2(x,e),$$

where a_7 is a small positive constant for given problem.

This Lemma can be proved using Lemma 2, Lemma 3 and Theorem 1 in similar fashion of the proofs of the previous lemmas. Lemma 4 provides theoretical guarantee for quality of solution of Eq. (2.21). Lemma 4 is useful to prove the bound of sub-selection version of Eq. (2.4):

$$\{\hat{\beta}, e_\Omega, \hat{\tau}\} = \arg\min_{\hat{\beta},e_\Omega,\hat{\tau}} \|y_\Omega \odot \hat{\tau} - U_\Omega\hat{\beta} - e_\Omega\|_2^2 + \gamma\|e_\Omega\|_1, \tag{2.22}$$

where U_Ω is the singular vectors of sub-selection dictionary $\Omega \odot A = A_\Omega = \Omega \odot USV^T$ and other terms have the same meaning as in the above discussion. The number

of rows of Eq. (2.21) and Eq. (2.22) are much smaller than that of Eq. (2.3) and Eq. (2.4), which can effectively reduce the computational complexity. An object classification algorithm based on TCSR is summarized in Algorithm 1.

Algorithm 1 Transform-invariant Collaborative Sub-representation (TCSR)

Input: Training data matrix A, query image y, and initial transformation τ_0 of y.

Generate l random selection operator $\Omega^1, \ldots, \Omega^l$

for $q = 1, \ldots, l$ **do**

$A_\Omega = A \odot \Omega^q$

Compute $y_\Omega = y \odot \Omega^q$

Set U_1 as the first η_1 column vectors of U_Ω where $A_\Omega = U_\Omega S V^T$

Solving Eq. (2.21) to get $\hat{x}, e_\Omega, \hat{\tau}^q$

Compute residue $r_i^q = \|y_\Omega \odot \hat{\tau}^q - A_\Omega \hat{x}_i\|$ for each class i

end for

Compute $\hat{\tau} = E[r_i]$

Compute $identity(y) = \arg\min_i E[r_i]$

Compute $transform(y) = E[\hat{r}^q]$

Output: $identity(y)$ and $transform(y)$

2.4 Experiments

In this section, we conducted extensive experiments to validate the acceleration performance of the proposed sub- representation based methods against transformation estima- tion, collaborative representation, and transform-invariant col- laborative representation, respectively. For fair comparisons, we download the code of these

algorithms from their websites and follow their default parameter settings. Three databases are used for training or testing: Multi-PIE [17], Extended Yale B(EYB) [18] and AR [19]. All experiments are conducted on a desktop computer with Intel iCore 7 3.4GHz CPU. Matlab version is 2012a.

2.4.1 Transformation Estimation

First, we test our sub-selection for transformation estima- tion algorithm (STE, i.e. Eq. (2.18)) using public database CMU Multi-PIE database [17]. The images of 100 subjects from Session 2 are chosen and all images are resized to 640×480. The areas of human faces are used as the region of interest (ROI) and an artificial transformation of x and y directions are introduced to the ROI. The reference face area is resized to 160×120. We compare STE with the hierarchical model- based motion estimation (HMME, i.e. Eq. (2.17)) [16] algorithm. Artificial translation in both x and y directions are added to the position of ROI. Suppose the groundtruth rectangle position is a 4-D vector R_1 consisting of x, y coordinates, width and height of the rectangle, while R_2 is a similar vector for the result rectangle. The accuracy of the estimation is calculated as: $acc = \|R_1 - R_2\|_2 / \|R_1\|2$ The translation is set as 10, 20, 30, 40 pixels respectively. We set the sample rate as 1/5 (i.e. $|\Omega| \approx n/5$). The result is shown in Table 2.1. In the table, the proposed algorithm is 2 to 3 times faster than the original HMME algorithm while the loss of estimation accuracies is no more than 0.1%. These experiments indicate that the sub-selection technique effectively reduces the computational complexity while preserving the accuracies.

2.4.2 Face Recognition

We use the proposed TCSR (i.e. Algorithm 1) for face recognition to validate its benefits. We compare the proposed TCSR with original collaborative representation

Translation		10	20	30	40
HMME [16]	Accucracy	99.6%	99.1%	94.9%	74.3%
	Time	0.19	0.38	0.60	0.43
Proposed	Accuracy	99.5%	99.1%	95.5%	75.3%
	Time	0.07	0.15	0.21	0.14

Table 2.1: Result Comparison of Transformation Estimation. Accuracy stands for accuracy and Time stands for average execution time for each image.

classification(CRC) [3] algorithm on the Multi-PIE [17], Extended Yale B(EYB) [18] and AR databases [19]. For fair comparison, we follow the experimental setting in [3]. The initial position of all the testing and training images are automatically detected by Viola and Jones face detector [20]. For the setting of Multi- PIE, the first 100 subjects in Session 1 with illuminations 0, 1, 7, 13, 14, 16, 18 are used as the training set while the first 100 subjects in Session 3 with illumination 3, 6, 11, 19 are used for testing. The face areas in the training images are all resized to 160×120. The sample rate is 1/15; For the setting of EYB, 20 subjects are selected. For each subject, 32 randomly selected frontal images are used for training, with 29 of the remaining images for testing. The face areas are resized to 192×168. The sample rate is 1/10; For the setting of AR, 100 subjects and 7 images per subject are selected as the training set, while 100 subjects and 6 images per subject for testing. There are 50 male faces and 50 female faces with various facial expressions and illuminations. The face areas are resized to 165×120. The sample rate is 1/8.

The experimental result is tabulated in Table 2.2. The recognition rate of proposed TCSR algorithm is almost the same as the original CRC [3] algorithm with difference less than 0.5while the speed is 3 to 10 times faster than the origin version related to the sample rate. Also the variation of illuminations of the training set and testing set affects the results. While the variation of the training set is sufficient to represent the query images, the sample rate can be lower, otherwise

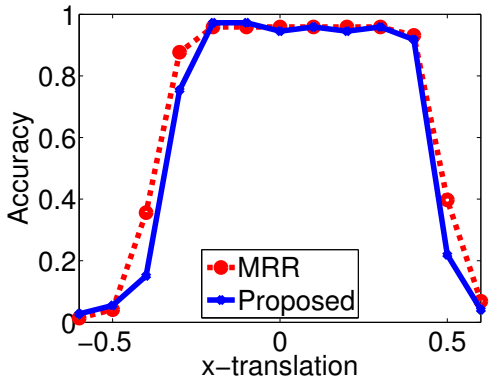
more data are needed. All these results have validated the proposed collaborative sub-representation has outperformed the collaborative representation in speed while achieving comparable accuracy.

		MPIE	EYB	AR
CRC [3]	Accuracy	88.7%	99.4%	86.1%
	Average Time	2.3	3.85	1.3
Proposed	Accuracy	89.1%	99.1%	86.6%
	Average Time	0.30	0.30	0.67

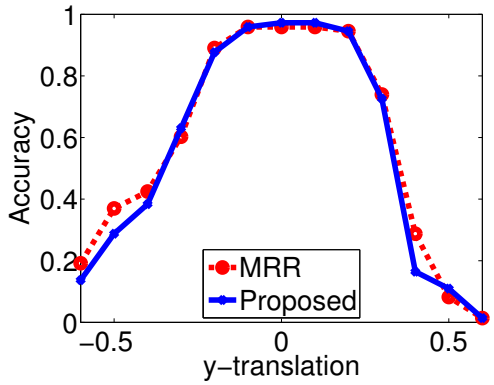
Table 2.2: Recognition rates and execution time.

2.4.3 Transformation Invariant Face Recognition

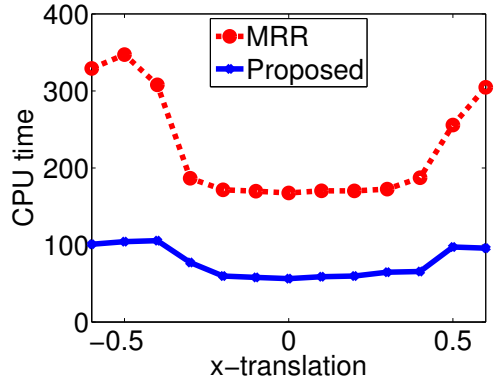
In this section, we conduct experiments to validate the benefit of TCSR under misalignment. First, we compare the proposed algorithm with state-of-the-art methods, MRR [4], RASR [11], TSR [10] on the Multi-PIE [17] database. The first 100 subjects in Session 1 with illuminations 0, 1, 7, 13, 14, 16, 18 are used as the training set while the first 100 subjects in Session 2 with illumination 10 are used for testing. Artificial transformation of 5 pixels translation in both x and y directions is added to the test images. The sample rate is set to 1/5. The recognition rates of RASR, TSR, MRR and the proposed algorithm are 91.8%, 89.1% 95.9% 95.9% respectively, while the average execution times are 95.9, 15.5, 4.9, 1.4 respectively. The RASR will try to fit all identities one by one, which makes it very slow in big training set. Although TSR is faster, its less accurate. MRR is significantly faster than RASR and TSR and also more accurate. In the following experiments we shall only compare our algorithm with MRR. Also note that in this experiment, our proposed algorithm is nearly 3 times faster than MRR while maintaining almost the same accuracy.



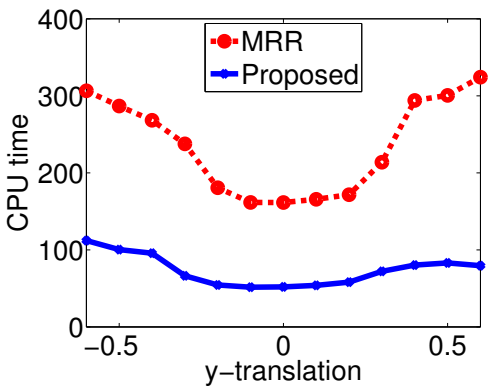
(a) Accuracy with only x direction translation.



(b) Accuracy with only y direction translation.



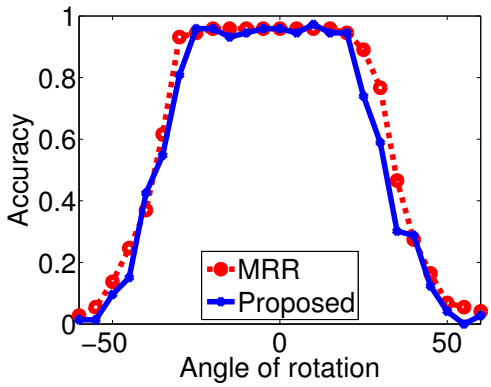
(c) Average Time with only x direction translation.



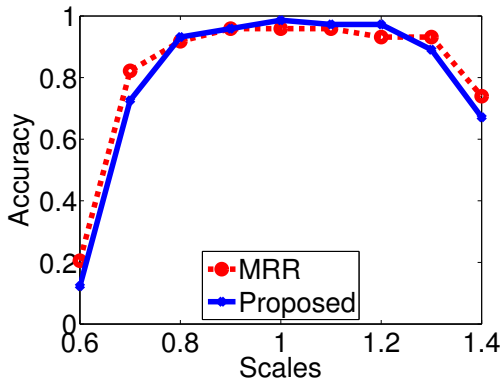
(d) Average Time with only y direction translation.

Figure 2.2: Recognition rates and Speed with Translation.

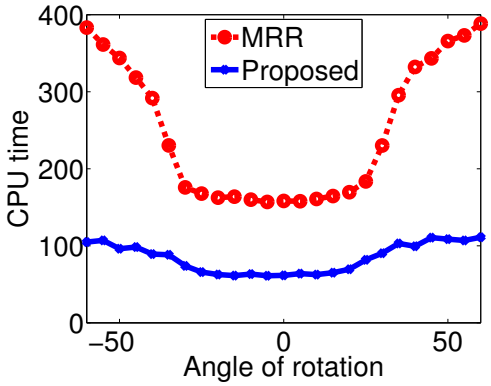
Next, we conduct experiments on the Multi-PIE database with various kinds of transformations. The experimental setting is the same as the previous experiment. Artificial transformations include x direction translation, y direction translation, in-place rotation and scales. These experiments compare the performance of MRR [4] and the proposed algorithm. Here we use the sample rate 1/5. The experimental results are shown in Fig. 2.2 and 2.3. The proposed algorithm is 3 to 4 times faster than the MRR, while the difference of their accuracy is less than 1%. Due to the redundancy of data, the sub-selection method preserve the accuracy very well.



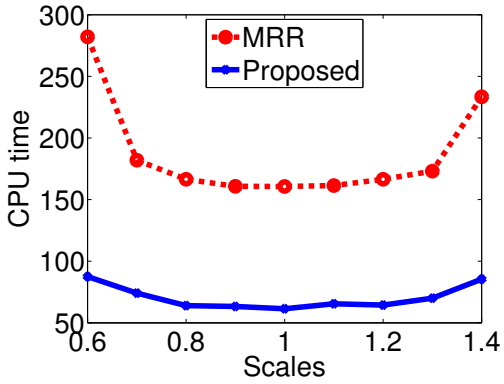
(a) Accuracy with only in-place rotation.



(b) Accuracy with only scale variantion.



(c) Average Time with only in-place rotation.



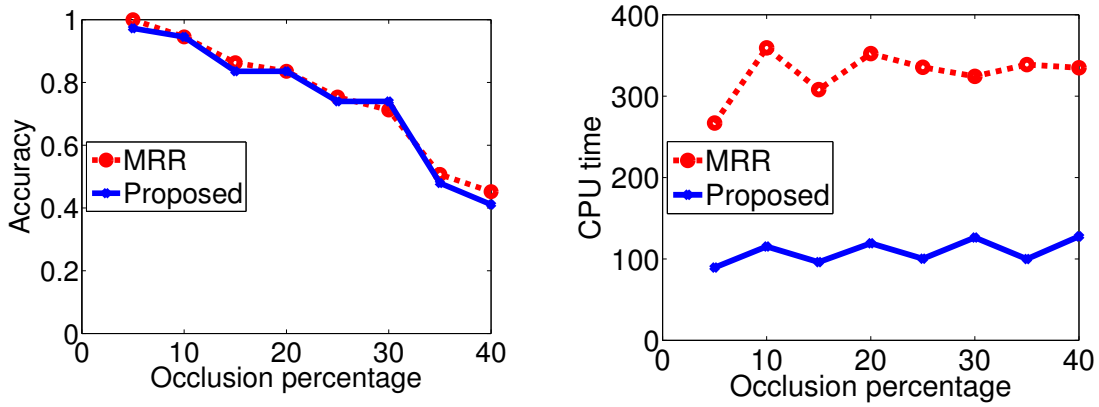
(d) Average Time with only scale variation.

Figure 2.3: Recognition rates and Speed with Scale and Rotation.

All these experiments validate that the proposed TCSR can handle various misalignments with much less computational cost than existing methods.

2.4.4 Recognition Despite Random Block Occlusion

In this section, we further validate the robustness of TCSR by comparing our method and MRR [4] on images with random block occlusions. This kind of experiment has been conducted in the MRR and RASR [11], so we follow the setting of the experiment and use the dataset as the same as those used in the previous experiment. The sample rate is 1/3. The training set and testing set are from the Multi-PIE



(a) The recognition rate(accuray) vs. occlusion percentage.

(b) The execution time vs. occlustion percentage.

Figure 2.4: Accuracy and CPU time.

database. The first 100 subjects in Session 1 are used as the training set. And the testing set is 100 subjects from Session 2. Various levels of block occlusion are added to the testing images. The testing results are shown in Fig. 2.4. TCSR has almost the same recognition rates as those of the MRR under various levels of block occlusions. However, the TCSR is 2 to 3 times faster than the MRR algorithm. These experiments validate the benefit of proposed TCSR in handling occlusions.

2.5 Conclusion

This chapter proposed a novel sub-representation method for image representation. We have theoretically proved its benefit for handling sparse big noise over previous meth1ods. Combining it with existing techniques, we proposed a transform-invariant sub-representation, which can efficiently handle misalignment, occlusion and big noise problems in practical applications. The benefit of proposed methods were not only theoretically proved but also empirically validated by extensive experiment results on practical applications.

In the future, we would like to extend our approach in several directions. First, it is worth investigating beyond the simple sparse regularization to group sparsity [21]. Second, we would like to apply our approach to more applications such as MR imaging reconstruction problem [22], [23].

CHAPTER 3

SUB-SELECTIVE QUANTIZATION FOR LARGE-SCALE IMAGE SEARCH

Performing hashing on floating-point features is a increasingly popular technique to handle large-scale high-dimensional data. This chapter proposed a novel unsupervised hashing approach based on sub-selection. This chapter also provides the theoretical guarantee for the performance of the proposed approach. It is shown that the proposed approach is able to achieve same level of accuracy with less than 1/10 of computational cost [12].

3.1 Introduction

Similarity search has stood as a fundamental technique used in many vision related applications including object recognition [24, 25], image retrieval [26, 27], image matching [28, 29], etc. The explosive growth of visual content on the Internet has made this task more challenging due to the high storage and computation overhead. To this end, mapping high-dimensional image descriptors to compact binary codes has been suggested, leading to considerable efficiency gains in both storage and similarity computation of images. The reason is simple: compact binary codes are much more efficient to store than floating-point feature vectors, and meanwhile similarity based on Hamming distances among binary bits is much easier to compute than Euclidean distances among real-valued features.

The benefits of binary encoding, also known as *Hashing* and *Quantization* in literature, have motivated a tremendous amount of research in binary code generation such as [30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 27, 40, 41, 42, 29, 43]. The main

challenge of these approaches is how to effectively incorporate domain knowledge into traditional models [21], and how to efficiently solve them [22]. The composite prior models are promising solutions because of their flexibility in modeling prior knowledge and their computational efficiency [22, 23]. Common in many methods, the first step has been adopted to leverage a linear mapping to project original features in high dimensions to lower dimensions. The representatives include Locality Sensitive Hashing (LSH) [30], Spectral Hashing (SH) [35], PCA Quantization (PCAQ) [27], Iterative Quantization (ITQ) [37], and Isotropic Hashing (IsoH) [44]. LSH uses random projections to form such a linear mapping, which is categorized into *data-independent* approaches since the used coding (hash) functions are fully independent of training data. Although learning-free, LSH requires long codes to achieve satisfactory accuracy. In contrast, *data-dependent* approaches can obtain high-quality compact codes by learning from training data. Specifically, PCAQ applies PCA to project the input data onto a low-dimensional subspace, and simply thresholds the projected data to generate binary bits each of which corresponds to a single PCA projection. Following PCAQ, SH, ITQ, and IsoH all employ PCA to acquire a low-dimensional data embedding, and then propose different postprocessing schemes to produce binary bits. A common drawback of the above learning-driven hashing methods is the expensive computational cost in matrix manipulations.

In this chapter, we demonstrate that the most time-consuming matrix operations encountered in code learning, typically data projection and rotation, can be performed in a more efficient manner. To this end, we propose a fast matrix multiplication algorithm using a sub-selection [6] technique to accelerate the learning of coding functions. Our algorithm is motivated by the observation that the degree of the algorithm parameters is usually very small compared to the number of entire data

samples. Therefore, we are able to determine these parameters merely using partial data samples.

The contributions of this chapter are three-folds: **(1)** To handle large-scale data, we propose a sub-selection based matrix multiplication algorithm and demonstrate its benefits theoretically. **(2)** We develop two fast quantization methods PCAQ-SS and ITQ-SS by combining the sub-selective algorithm with PCAQ and ITQ. **(3)** Extensive experiments are conducted to validate the efficiency and effectiveness of the proposed PCAQ-SS and ITQ-SS, which indicate that ITQ-SS can achieve an up to 30 times acceleration of binary code learning yet with an imperceptible loss of accuracy.

3.2 Background and Related Work

Before describing our methods, we will briefly introduce the binary code learning problem and two popular approaches.

Binary Encoding is trying to seek a coding function which maps a feature vector to short binary bits. Let $X \in \mathbb{R}^{n\times d}$ be the matrix of input data samples, and the i-th data sample $x_i \in \mathbb{R}^{1\times d}$ be the i-th row in X. Additional, X is made to be zero-centered. The goal is then to learn a binary code matrix $B \in \{-1, 1\}^{n\times c}$, where c denotes the code length. The coding functions of several hashing and quantization methods can be formulated into $h_k(x) = sgn(xp_k)$ $(k = 1, \ldots, c)$, where $p_k \in \mathbb{R}^d$ and the sign function $sgn(\cdot)$ is defined as:

$$sgn(v) = \begin{cases} 1, & \text{if } v > 0; \\ -1, & \text{otherwise.} \end{cases}.$$

Hence, the coding process can be written as $B = sgn(XP)$, where $P = [p_1, \cdots, p_c] \in R^{d\times c}$ is the projection matrix.

PCA Quantization (PCAQ) [27] finds a linear transformation $P = W$ that maximizes the variance of each bit and makes the c bits mutually uncorrelated. W is obtained by running Principal Components Analysis (PCA). Let $[W, \Lambda] = eig(\cdot, c)$ be a function which returns the first c eigenvalues in a diagonal matrix $S \in \mathbb{R}^{c \times c}$ and the corresponding eigenvectors as columns of $W \in \mathbb{R}^{d \times c}$. The whole procedure is summarized in Algorithm 2. While it is not a good coding method, its PCA step has widely used as an initial step of many sophisticated coding methods. However, the computation of PCA involves a multiplication with high-dimensional matrix X, which consumes considerable amount of memory and computation time. We will address the efficiency issue of PCAQ in the next section.

Algorithm 2 PCA Quantization (PCAQ)

1: **Input:** Zero-centered data $X \in \mathbb{R}^{n \times d}$, code length c.

2: **Output:** $B \in \{-1, 1\}^{n \times c}$, $W \in \mathbb{R}^{d \times c}$.

3: $cov = X^T X$;

4: $[W, \Lambda] = eig(cov, c)$;

5: $B = sgn(XW)$.

Iterative Quantization (ITQ) [37] improves the quality of PCAQ by iteratively finding the optimal rotation matrix R on the projected data to minimize the quantization error. This is done through finding an appropriate orthogonal rotation by minimizing:

$$Q(B, R) = \|B - VR\|_F^2, \tag{3.1}$$

where $V = XW$ is the PCA projected data. This equation is minimized using the spectral-clustering like iterative quantization procedure [45]. The whole procedure is summarized in Algorithm 3, where $svd(\cdot)$ indicates singular value decomposition.

The ITQ method converges in a small number of iterations and is able to achieve high-quality binary codes compared with state-of-the-art coding methods. However, it involves not only multiplications with high-dimensional matrices (*e.g.*, $X^T X$ and $B^T V$) in the PCA step, but also those inside each quantization iteration, which makes it very slow in training. In the next section, we will propose a method to overcome this drawback while preserving almost the same level of coding quality.

Algorithm 3 Iterative Quantization (ITQ)

Input: Zero-centered data $X \in \mathbb{R}^{n \times d}$, code length c, iteration number N.
Output: $B \in \{-1, 1\}^{n \times c}$, $W \in \mathbb{R}^{d \times c}$.
$cov = X^T X$;
$[W, \Lambda] = eig(cov, c)$;
$V = XW$;
initialize R as an Orthogonal Gaussian Random matrix;
for $k = 1$ **to** N **do**
    $B = sgn(VR)$;
    $[S, \Lambda, \hat{S}] = svd(B^T V)$;
    $R = \hat{S} S^T$;
end for
$B = sgn(VR)$.

3.3 Methodology

According to our previous discussion, the common bottleneck of many existing methods is high dimensional matrix multiplication. However, dimensions of the product of these multiplication is relatively small. This motivates us to search for

good approximation of those products using a subset of data, which results in our sub-selective matrix multiplication approach.

3.3.1 Sub-selective Matrix Multiplication

The motivation behind sub-selective multiplication can be explain intuitively using data distribution. First of all, the data matrix X is low-rank compared to n when $d << n$. Hence, all samples can be linear represented by a small subset of all. In previous discussion, the quantization algorithms try to learn the parameters, i.e. W and R, that can transform data distribution according to specific criteria (e.g. variances). If data are distributed closely to uniform, then a sufficient random subset can represent the full set well enough. Therefore we can find those parameters by solving the optimization problems in the selected subsets.

We begin with introduction to the notations of sub-selection. Let $\Omega \subset \{1, \ldots, n\}$ denotes the indexes of selected rows of matrix ordered lexicographically and $|\Omega| = m$ denotes the cardinality of Ω. With the same notations as previous section, the sub-selection operation on X can be expressed as $X_\Omega \in \mathbb{R}^{m \times d}$ that consists of row subset of X. For easy understanding we can consider X_Ω as $I_\Omega X$ where X multiply by a matrix $I_\Omega \in \{0, 1\}^{m \times n}$ that consists of random row subset of the identify matrix I_n.

With sub-selection operation, for matrix $Y \in \mathbb{R}^{n \times d_1}$ and $Z \in \mathbb{R}^{n \times d_2}$, where $d_1, d_2 \ll n$, sub-selective multiplication use $\frac{n}{m} Y_\Omega^T Z_\Omega$ to approximate $Y^T Z$. And for a special case $Y^T Y$, its sub-selection approximation is $\frac{n}{m} Y_\Omega^T Y_\Omega$. The complexity of multiplication is now reduced from $O(nd_1d_2)$ to $O(md_1d_2)$. Before we apply this methods to binary quantization, we will first examine if it's theoretically sound.

We will prove a bound for sub-selective multiplication. Before providing our analysis, we first introduce a key result (Lemma 5 below) that will be crucial for the later analysis.

Lemma 5. *(McDiarmid's Inequality [15]): Let $X_1, ..., X_n$ be independent random variables, and assume f is a function for which there exist t_i, $i = 1, ..., n$ satisfying*

$$\sup_{x_1,...,x_n,\hat{x}_i} |f(x_1, ..., x_n) - f(x_1, ..., \hat{x}_i, ..., x_n)| \leq t_i \tag{3.2}$$

where $\hat{x}_i$ indicates replacing the sample value x_i with any other of its possible values. Call $f(X_1, ..., X_n) := Y$. Then for any $\epsilon > 0$,

$$\mathbb{P}\left[Y \geq \mathbb{E}[Y] + \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n t_i^2}\right) \tag{3.3}$$

$$\mathbb{P}\left[Y \leq \mathbb{E}[Y] - \epsilon\right] \leq exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n t_i^2}\right) \tag{3.4}$$

Let U be an $n \times r$ matrix whose columns span the r-dimensional subspace S. Let $P_S = U(U^TU)^{-1}U^T$ denotes the projection operator onto S. The "*coherence*" [14] of U is defined to be

$$\mu(S) := \frac{n}{r} \max_j \|P_S e_j\|_2^2, \tag{3.5}$$

where e_j represents a standard basis element. $\mu(S)$ measure the maximum magnitude attainable by projecting a standard basis element onto S. Note that $1 \leq \mu(S) \leq \frac{n}{r}$. Let $z = [\|U_1\|_2, \ldots, \|U_i\|_2, \ldots, \|U_n\|_2]^T \in \mathbb{R}^n$, where each element of z is l_2-norm of one row in U. Thus, based on "*coherence*", we define "*row coherence*" to be the quantity

$$\phi(S) := \mu(z). \tag{3.6}$$

By plugging in the definition, we have $\phi(S) = \frac{n\|U\|_{2,\infty}^2}{\|U\|_F^2}$, where $\|\cdot\|_{2,\infty}$ means first compute the l_2-norm of each row then compute l_∞-norm of result vector.

The key contribution of this chapter is the following two theorems that form the analysis of bounds to sub-selective matrix multiplication. We start from the special case $Y_\Omega^T Y_\Omega$.

Theorem 3. *: Suppose $\delta > 0$, $Y \in \mathbb{R}^{n\times d}$ and $|\Omega| = m$, then*

$$(1-\alpha_1)\frac{m}{n}\|Y\|_F^2 \leq \|Y_\Omega\|_F^2 \leq (1+\alpha_1)\frac{m}{n}\|Y\|_F^2 \tag{3.7}$$

with probability at least $1-2\delta$, where $\alpha_1 = \sqrt{\frac{2\phi_1(Y)^2}{m}\log(\frac{1}{\delta})}$ and $\phi_1(Y) = \frac{n\|Y\|_{2,\infty}^2}{\|Y\|_F^2}$.

Proof. We use McDiarmid's inequality from Lemma 5 for the function $f(X_1,\ldots,X_m) = \sum_{i=1}^m X_i$ to prove this. Set $X_i = \sum_{j=1}^d |Y_{\Omega(i),j}|^2$. Let $\|\cdot\|_1$ denotes the l_1 norm of matrix. Since $\sum_{j=1}^d |Y_{\Omega(i),j}|^2 \leq \|Y\|_{2,\infty}^2$ for all i, we have

$$\left|\sum_{i=1}^m X_i - \sum_{i\neq k} X_i - \hat{X}_k\right| = \left|X_k - \hat{X}_k\right| \leq 2\|Y\|_{2,\infty}^2. \tag{3.8}$$

We first calculate $\mathbb{E}[\sum_{i=1}^m X_i]$ as follows. Define $\mathbb{I}_{\{\}}$ to be the indicator function, and assume that the samples are taken uniformly with replacement.

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^m X_i\right] &= \mathbb{E}\left[\sum_{i=1}^m\sum_{j=1}^d |Y_{\Omega(i,j)}|^2\right] \\ &= \sum_{i=1}^m\left[\mathbb{E}\left[\sum_{k=1}^n\sum_{j=1}^d |Y_{k,j}|^2\mathbb{I}_{\{\Omega(i)=k\}}\right]\right] = \frac{m}{n}\|Y\|_F^2.\end{aligned} \tag{3.9}$$

Invoking the Lemma 5, the left hand side is

$$\mathbb{P}\left[\sum_{i=1}^m X_i \leq \mathbb{E}\left[\sum_{i=1}^m X_i\right] - \epsilon\right] = \mathbb{P}\left[\sum_{i=1}^m X_i \leq \frac{m}{n}\|Y\|_F^2 - \epsilon\right]. \tag{3.10}$$

We can let $\epsilon = \alpha\frac{m}{n}\|y\|_F^2$ and then have that this probability is bounded by

$$\exp\left(\frac{-2\alpha^2(\frac{m}{n})^2\|Y\|_F^4}{4m\|Y\|_{2,\infty}^4}\right) \tag{3.11}$$

Thus, the resulting probability bound is

$$\mathbb{P}\left[\|Y_\Omega\|_F^2 \geq (1-\alpha)\frac{m}{n}\|Y\|_F^2\right] \geq 1 - \exp\left(\frac{-\alpha^2 m\|Y\|_F^4}{2n^2\|Y\|_{2,\infty}^4}\right). \tag{3.12}$$

Substituting our definitions of $\phi_1(Y) = \frac{n\|Y\|_{2,\infty}^2}{\|Y\|_F^2}$ and $\alpha_1 = \sqrt{\frac{2\phi_1(Y)^2}{m}\log(\frac{1}{\delta})}$ shows that the lower bound holds with probability at least $1-\delta$. The argument for the upper bound can be proved similarly. The Theorem now follows by applying the union bound. □

Now we analysis the property of general case $Y_\Omega^T Z$.

Theorem 4. *: Suppose $\delta > 0$, $Y \in \mathbb{R}^{n\times d_1}$, $Z \in \mathbb{R}^{n\times d_2}$ and $|\Omega| = m$, then*

$$(1-\beta_1)^2(\frac{m}{n})^2\|Y^TZ\|_F^2 \leq \|Y_\Omega^T Z_\Omega\|_F^2 \leq (1+\beta_2)^2\frac{m}{n}\|Y^TZ\|_F^2 \tag{3.13}$$

with probability at least $1-2\delta$, where

$$\beta_1 = \sqrt{\frac{2nd_1d_2\mu(S_Y)\mu(S_Z)}{m^2\|Y^TZ\|_F^2}\log(\frac{1}{\delta})}$$

and

$$\beta_2 = \sqrt{\frac{2d_1d_2\mu(S_Y)\mu(S_Z)}{m\|Y^TZ\|_F^2}\log(\frac{1}{\delta})}$$

.

Proof. This theorem can be proved by involving McDiarmid's inequality in similar fashion to the proof of Theorem 3. Let $X_i = Y_{\Omega(i)}^T Z_{\Omega(i)} \in \mathbb{R}^{d_1\times d_2}$, where $\Omega(i)$ denotes the i^{th} sample index, $Y_{\Omega(i)} \in \mathbb{R}^{d_1\times 1}$ and $Z_{\Omega(i)} \in \mathbb{R}^{d_2\times 1}$.

Let our function $f(X_1,\ldots,X_m) = \|\sum_{i=1}^m X_i\|_F = \|Y_\Omega^T Z_\Omega\|_F$. First, we need to bound $\|X_i\|$ for all i. Observe that $\|Y_{\Omega(i)}\|_F = \|Y^Te_i\|_2 = \|P_{S_Y}e_i\|_2 \leq \sqrt{d_1\mu(S_Y)/n}$ by assumption, where S_Y refers to the subspace span by Y. Likewise, we have $\|Z_{\Omega(i)}\|_F \leq \sqrt{d_2\mu(S_Z)/n}$, where S_Z refers to the subspace span by Z. Thus,

$$\begin{aligned}\|X_i\|_F = \|Y_{\Omega(i)}^T Z_{\Omega(i)}\|_F &\leq \|Y_{\Omega(i)}\|_F\|Z_{\Omega(i)}\|_F \\ &\leq \sqrt{d_1d_2\mu(S_Y)\mu(S_Z)/n^2}.\end{aligned} \tag{3.14}$$

Then $|f(X_1,\ldots,X_m) - f(X_1,\ldots,\hat{X}_K,\ldots,X_m)|$ is bounded by

$$\begin{aligned}
&\left| \left\| \sum_{i=1}^{m} X_i \right\|_F - \left\| \sum_{i\neq k} X_i + \hat{X}_k \right\|_F \right| \\
\leq & \|X_k - \hat{X}_k\|_F \leq \|X_k\|_F + \|\hat{X}_k\|_F \\
\leq & 2\sqrt{d_1 d_2 \mu(S_Y)\mu(S_Z)/n^2}, \qquad (3.15)
\end{aligned}$$

where the first two inequalities follow from the triangular inequality. Next we calculate the bound for $\mathbb{E}\left[f(X_1,\ldots,X_m)\right] = \mathbb{E}[\|\sum_{i=1}^{m} X_i\|_F]$. Assume again that the samples are taken uniformly with replacement.

$$\begin{aligned}
\mathbb{E}\left[\left\| \sum_{i=1}^{m} X_i \right\|_F^2 \right] &= \mathbb{E}\left[\left\| \sum_{i=1}^{m} Y_{\Omega(i)}^T Z_{\Omega(i)} \right\|_F^2 \right] \\
&= \sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2} \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{k_1,j}^2 Z_{k_2,j}^2 \mathbb{I}_{\{\Omega(i)=j\}} \right] \qquad (3.16) \\
&= \sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2} m \sum_{j=1}^{n} Y_{k_1,j}^2 Z_{k_2,j}^2 \frac{1}{n} = \frac{m}{n}\|Y^T Z\|_F^2 \qquad (3.17)
\end{aligned}$$

The step (3.16) follows because of our assumption that sampling if uniform with replacement.

Since $\mathbb{E}[\|\sum_{i=1}^{m} X_i\|_F] \leq \mathbb{E}[\|\sum_{i=1}^{m} X_i\|_F^2]^{1/2}$ by Jensen's inequality, we have

$$\mathbb{E}\|\sum_{i=1}^{m} X_i\|_F] \leq \sqrt{\frac{m}{n}}\|Y^T Z\|_F.$$

Using Jensen's inequality and indicator function in similar fashion, we also have bound for the left side:

$$\begin{aligned}
&\mathbb{E}\left[\|\sum_{i=1}^{m} X_i\|_F\right] \geq \left\|\sum_{i=1}^{m} \mathbb{E}\left[X_i\right]\right\|_F \\
&= \left\|\sum_{i=1}^{m} \mathbb{E}\left[Y_{\Omega(i)}^T Z_{\Omega(i)}\right]\right\|_F \\
&= \left\|\sum_{i=1}^{m} \mathbb{E}\left[\sum_{j=1}^{n} Y_j^T Z_j \mathbb{I}_{\{\Omega(i)=j\}}\right]\right\|_F \\
&= \|\frac{m}{n} Y^T Z\|_F = \frac{m}{n}\|Y^T Z\|_F
\end{aligned} \tag{3.18}$$

Letting $\epsilon_1 = \beta_1 \frac{m}{n}\|Y^T Z\|_F$ and plugging into Equation (3.4), we then have that probability is bounded by

$$\exp\left(\frac{-2\beta_1^2(\frac{m}{n})^2\|Y^T Z\|_F^2}{4nd_1d_2\mu(S_Y)\mu(S_Z)/n^2}.\right) \tag{3.19}$$

Thus, the resulting probability bound is

$$\begin{aligned}
&\mathbb{P}\left[\|Y_\Omega^T Z_\Omega\|_F^2 \geq (1-\beta_2)^2(\frac{m}{n})^2\|Y^T Z\|_F^2\right] \\
&\qquad \geq 1 - \exp\left(\frac{-\beta_1^2 m^2\|Y^T Z\|_F^2}{2nd_1d_2\mu(S_Y)\mu(S_Z)}\right).
\end{aligned} \tag{3.20}$$

Substituting our definitions of $\mu(S_Y)$, $\mu(S_Z)$ and β_1 shows that the lower bound holds with probability at least $1-\delta$.

Letting $\epsilon_2 = \beta_2\sqrt{\frac{m}{n}}\|Y^T Z\|_F$ and plugging into Equation (3.3), we then have that probability is bounded by:

$$\exp\left(\frac{-2\beta_2^2(\frac{m}{n})\|Y^T Z\|_F^2}{4nd_1d_2\mu(S_Y)\mu(S_Z)/n^2}\right) \tag{3.21}$$

Thus, the resulting probability bound is

$$\begin{aligned}
&\mathbb{P}\left[\|Y_\Omega^T Z_\Omega\|_F^2 \leq (1+\beta_2)^2\frac{m}{n}\|Y^T Z\|_F^2\right] \\
&\qquad \geq 1 - \exp\left(\frac{-\beta_2^2 m\|Y^T Z\|_F^2}{2d_1d_2\mu(S_Y)\mu(S_Z)}\right)
\end{aligned} \tag{3.22}$$

Substituting our definitions of $\mu(S_Y)$, $\mu(S_Z)$ and β_2 shows that the upper bound holds with probability at least $1-\delta$. The theorem now follows by applying the union bound, completing the proof. □

The above two theorems prove that the product of sub-selective multiplication will be very close the original product of full data with high probability.

3.3.2 Case Studies: Sub-selective Quantization

With the theoretical guarantee, we are now ready to apply sub-selective multiplication on existing quantization methods, i.e. PCAQ [27], ITQ [37]. A common initial step of them is PCA projection (e.g. Alg. 2 and Alg. 3). The time complexity for matrix multiplication X^TX is $O(nd^2))$ when $d < n$. For large n, this step could take up considerable amount of time. Hence, we can approximate it by $\frac{1}{m}X_\Omega^T X_\Omega$, which is surprisingly the covariance matrix of the selected samples. From statistics point of view, this could be intuitively interpreted as using the variance matrix of a random subset of samples to approximate the covariance matrix of full ones when the data is redundant. Now the time complexity is only $O(md^2)$, where $m \ll n$ in large dataset. For ITQ, the learning process includes dozens of iterations to find rotation matrix R (Alg. 3 line 7 to 11). We approximate R with $\hat{R} = S_r S_l$, where $S_l \Lambda S_r = B_\Omega^T V_\Omega$ is the SVD of $B_\Omega^T V_\Omega$, B_Ω and V_Ω are sub-selection version of B and V in Alg. 3 respectively. The time complexity of compute R is reduced from $O(nc^2)$ to $O(mc^2)$.

By replacing corresponding steps in original methods, we get two Sub-selective Quantization methods corresponding to PCAQ and ITQ, which are named PCAQ-SS, ITQ-SS. ITQ-SS is summarized in Algorithm 4. PCAQ-SS is the same as first 5 lines in Algorithm 4 plus one encoding step $B = sgn(V)$. It's omitted because of the page

limits. Complexity of original ITQ is $O(nd^2 + (p+1)nc^2)$. In contrast, complexity of ITQ-SS is reduced to $O(md^2 + pmc^2 + nc^2)$. The acceleration can be seen more clearly in the experimental results in the next section.

Algorithm 4 ITQ with Sub-Selection (ITQ-SS)

Input: Zero-centered data $X \in \mathbb{R}^{n \times d}$, code length c, iteration number p.

Output: $B \in \{-1, 1\}^{n \times c}$, $W \in \mathbb{R}^{d \times c}$.

1. Uniformly randomly generate $\Omega \subset [1:n]$;
2. $X_\Omega = \Omega \odot X$;
3. $cov = X_\Omega^T X_\Omega$;
4. $[W, \Lambda] = eig(cov, c)$;
5. $V = XW$;
6. initialize R as an Othorgonal Gaussian Random matrix;

for $k = 1$ **to** p **do**

uniformly randomly generate $\Omega \subset [1:n]$;

compute V_Ω;

$B_\Omega = sgn(V_\Omega R)$;

$[S, \Lambda, \hat{S}] = svd(B_\Omega^T V_\Omega)$;

$R = \hat{S} S^T$;

end for

7. $B = sgn(VR)$.

3.4 Evaluations

3.4.1 Experimental Setting

In this section, we evaluate the Sub-selective Quantization approaches on three public datasets: **CIFAR** [46] [1], **MNIST**[2] and **Tiny-1M** [27].

- **CIFAR** consists of 60K 32×32 color images that have been manually labelled to ten categories. Each category contains 6K samples. Each image in CIFAR is assigned to one mutually exclusive class label and represented by a 512-dimensional GIST feature vector [47].
- **MNIST** consists of 70K samples of 784-dimensional feature vector associated with digits from '0' to '9'. The true neighbours are defined semantic neighbours based on the associated digit labels.
- **Tiny-1M** consists of one million images. Each image is represented by a 384-dimensional GIST vector. Since manually labels are not available on Tiny-1M, Euclidean neighbours are computed and used as ground truth of nearest neighbour search.

We compare proposed methods **PCAQ-SS** and **ITQ-SS** with their corresponding unaccelerated methods **PCAQ** [27] and **ITQ** [37]. We also compare our methods to two baseline methods that follow similar quantization scheme $B = sgn(X\tilde{W})$: **1) LSH** [30], $\tilde{W}$ is a Gaussian random matrix; **2) SH** [35], which is based on quantizing the values of analytical eigenfunctions computed along PCA directions of the data. All the compared codes are provided by the authors.

Two types of evaluation are conducted following [37]. First, semantic consistency of codes is evaluated for different methods while class labels are used as ground truth. We report four measures, the **average precision of top 100 ranked images**

[1] http://www.cs.toronto.edu/ kriz/cifar.html

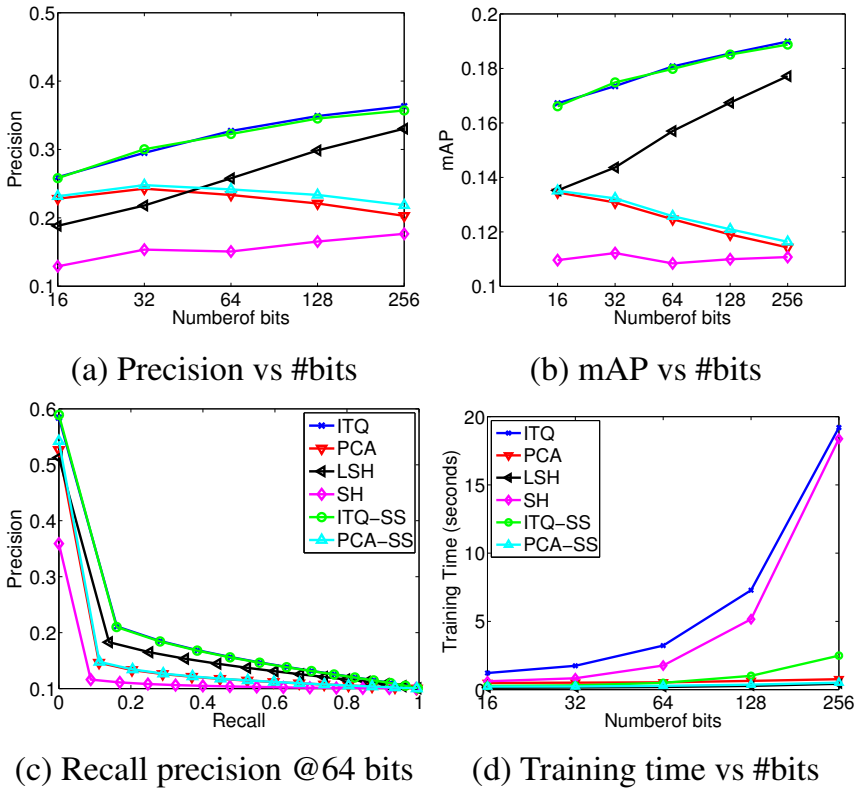
[2] http://yann.lecun.com/exdb/mnist/

for each query, **mean average precision**, **recall-precision curve** and **training time**, in CIFAR and MNIST. Second, we use the generated codes for nearest neighbour search, where Euclidean neighbours are used as ground truth. This experiment is conducted on Tiny-1M dataset. We report the three measures: **average precision of top 5% ranked images** for each query and **training time**. For both types of evaluation, the query algorithm and corresponding structure of binary code are the same, so **testing time** are exactly the same for all the methods except SH. Hence, it's omitted from the results. For the limit of page length, only parts of results are presented while the rest are put in the supplementary materials. All our experiments were conducted on a desktop computer with a 3.4GHz Intel Core i7 and 12GB RAM.

3.4.2 Results on CIFAR Dataset

The CIFAR dataset is partitioned into two parts: 59K images as a training set and 1K images as a test query set evenly sampled from ten classes. We uniformly randomly generate our sub-selective matrix Ω with cardinality equals to 1/40 of number of data points, i.e. $|\Omega| = m = n/40$.

Figure 3.1(a) and Figure 3.1(b) show complete precision of top 100 ranked images and mean average precision (mAP) over 1K query images for different number of bits. Figure 3.1(c) shows recall-precision curse of 64 bits code. For these three metrics, ITQ and ITQ-SS have the best performance. Both sub-selective methods (PCAQ-SS and ITQ-SS) preserve the performance of original methods (i.e. PCAQ and ITQ). Our results indicate that sub-selection preserve semantic consistency of original coding method. Figure 3.1(d) shows the training time of the two methods. Our method is about **4 to 8 times faster** than ITQ [37]. Original ITQ is the slowest among all the comparing methods, while the speed of the accelerated version ITQ-SS is comparable, if not superior, to the fastest methods. This is due to ITQ-SS reduce



(a) Precision vs #bits (b) mAP vs #bits

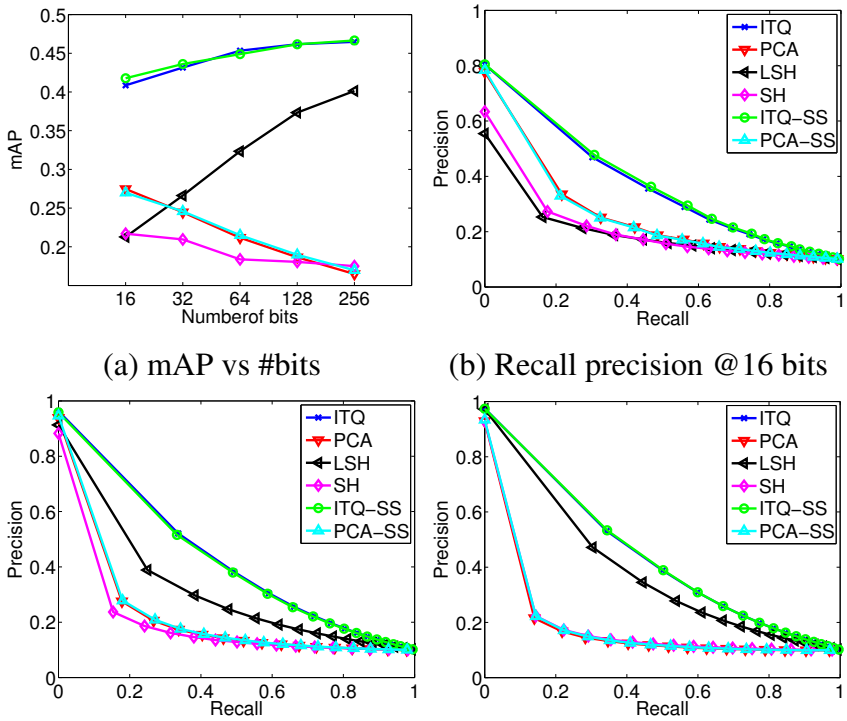
(c) Recall precision @64 bits (d) Training time vs #bits

Figure 3.1: The results on CIFAR. All the subfigures share the same legends.

the dimension of the problem from a function of n to that of m, where $m \ll n$. These results validate the benefits of sub-selection to preserve the performance of original method with far less training cost.

3.4.3 Results on MNIST Dataset

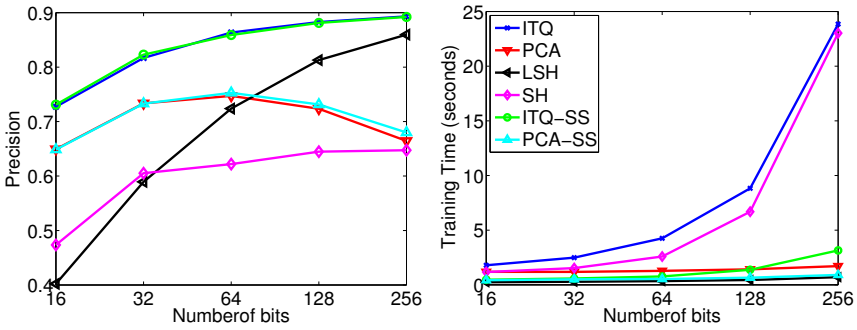
The MNIST dataset is splited into two subsets: 69K samples as a training set and 1K samples as a query set. While CIFAR dataset evaluates the performance of sub-selective quantization on complex visual features, MNIST evaluates that on raw pixel features. Similar to the previous experiment on CIFAR, we uniformly randomly generate our sub-selective matrix Ω with cardinality equals to $1/40$ of number of datapoints, i.e. $|\Omega| = m = n/40$. Figure 3.2(b) to Figure 3.2(d) shows three recall-



(a) mAP vs #bits (b) Recall precision @16 bits

(c) Recall precision @64 bits (d) Recall precision @256 bits

Figure 3.2: Results on MNIST. All the subfigures share the same set of legends.



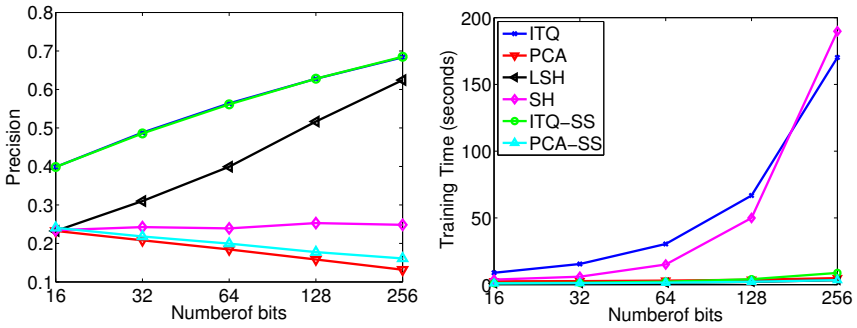
(a) Precision vs #bits (b) Training time vs #bits

Figure 3.3: The results on MNIST. All the subfigures share the same set of legends.

precision curves of Hamming ranking over 1K images corresponding to 16, 64 and 256 bits code. In all cases, the two curves of ITQ and proposed ITQ-SS are almost overlapping in all segments. Same trend can be seen for PCAQ and PCAQ-SS. Figure 3.2(a) and Figure 3.3(a) show complete precision of top 100 ranked images and mean

average precision (mAP) over 1K query images for different number of bits. The difference between ITQ and proposed ITQ-SS are almost negligible. The results confirm the trends seen in Figure 3.3(a). Figure 3.3(b) shows the training time of the two methods. Our method is about **3 to 8 times faster** than ITQ. The results of performance and training time are consistent with results on CIFAR. These results again validate the benefits of sub-selection.

3.4.4 Results on Tiny-1M Dataset



(a) Precision vs #bits. (b) Training time vs #bits.

Figure 3.4: The results on Tiny-1M. All the subfigures share the same set of legends.

For experiment without labelled groundtruth, a separate subset of 2K images of 80 million images are used as the test set while another one million images are used as the training set. We uniformly randomly generate our sub-selective matrix Ω with cardinality equals to 1/1000 of number of data points, i.e. $|\Omega| = m = n/1000$. Figure 3.4(a) shows complete precision of top 5% ranked images and mean average precision (mAP) over 1K query images for different number of bits. The difference between sub-selective methods (i.e. PCAQ-SS, ITQ-SS) and their counterparts (i.e. PCAQ, ITQ) are less than 1%. Figure 3.4(b) shows the training time of the two

methods. The ITQ-SS have achieved even bigger speed advantage, which is about **10 to 30 times faster** than ITQ. This is because the larger dataset samples are more redundant, making it possible to use smaller portion of data.

3.5 Discussion and Conclusion

All of the experimental results have verified the benefits of the sub-selective quantization technique whose parameters can be automatically learned from a subset of the original dataset. The proposed PCAQ-SS and ITQ-SS methods have achieved almost the same quantization quality as PCAQ and ITQ with only a small portion of training time. The advantage in training time is more prominent on larger datasets, *e.g.*, 10 to 30 times faster on Tiny-1M. Hence, for larger datasets good quantization quality can be achieved with an even lower sampling ratio.

One may notice that the speed-up ratio is not as same as the sampling ratio. This is because the training process of quantization includes not only finding the coding parameters but also generating the binary codes of the input dataset. The latter inevitably involves the operations upon the whole dataset, which costs a considerable number of matrix multiplications. In fact, this is one single step requiring matrix multiplications, thus enabling an easy acceleration by using parallel or distributed computing techniques. We will leave this problem to future work.

We accredit the success of the proposed sub-selective quantization technique to the effective use of sub-selection in accelerating the quantization optimization that involves large-scale matrix multiplications. Moreover, the benefits of sub-selection were theoretically demonstrated. As a case study of sub-selective quantization, we found that ITQ-SS can accomplish the same level of coding quality with significantly reduced training time in contrast to the existing methods. The extensive image

retrieval results on large image corpora with size up to one million further empirically verified the speed gain of sub-selective quantization.

CHAPTER 4

LARGE-SCALE MULTI-VIEW SPECTRAL CLUSTERING WITH BIPARTITE GRAPH

This chapter investigates the problem of unsupervised learning on large-scale multi-view data. We first proposed an efficient algorithm for constructing similarity graph for multi-view data. Then, we show that the constructed graph also significantly accelerate the clustering process. Extensive experiments on various data set demonstrate the proposed algorithm have achieve up to up to thousands times of acceleration [48].

4.1 Introduction

Clustering multi-view data is an important problem. In many real-world datasets, data are naturally represented by different features or views. This is due to the fact that data may be collected from different sources or be represented by different kind of features for different tasks. For example, documents can be written in different languages; gene can be measured by different techniques, e.g. gene expression, Single-nucleotide polymorphism (SNP), methylation; images can be described by different features like Gabor [47], HoG [49], GIST [47], LBP [50]. Different features capture different aspects of data and can be complementary to each other. Therefore, it is critical for learning algorithm to integrate these heterogeneous features to improve its accuracy and robustness. In this chapter, we focus on one specific unsupervised learning task, i.e., multi-view spectral clustering.

Recently, spectral clustering (SC) is drawing more and more attention because of its effectiveness [51, 52, 53, 54, 55, 56, 57]. However, the growth of the scale of data has rendered the multi-view clustering problem more challenging. None of the existing methods is applicable on large-scale multi-view data. In general, SC methods usually involve two time consuming steps. The first step is to construct the affinity graph and the second step is to compute the eigen-decomposition. The first step usually takes $O(n^2d)$ time while the second step takes $O(Kn^2)$ time, where n is the number of data points, d is the dimension of features and K is the number of clusters. Many works have been proposed to accelerate SC algorithm [58], [59], [60], [61], [62], [63]. These methods reply on various off-the-shelf projection or sampling methods [64, **?**, 6] to reduce the complexity of graph construction or eigen-decomposition. However, they only discuss the situation of handling single view data, which limits their usage. There are also SC methods that deal with multi-view data, such as [65], [66]. These methods try to model the multi-view clustering problem as solving local and global optimization among different views. Although they have achieved better accuracy than single-view SC methods, they are more computationally expensive due to the fact that they require iterations to reach consensus of different views or large-scale matrix inversion.

Another drawback of SC methods is that they usually do not provide natural extension to handle the out-of-sample problem [67, 68]. To address this problem, several methods have been proposed, e.g. [69, 58, 70, 68, 67]. They either rely on approximation of eigenfuncions [58, 68] or data projection such as error correcting output code (ECOC) method [71, 69] or regression model [67]. None of them address the out-of-sample problem in setting involved heterogeneous features.

In this chapter, we proposed a multi-view spectral clustering method that is able to deal with large-scale data. Our method is inspired by the large-scale semi-

supervised learning algorithm proposed in [72]. First, we generate consensus m salient points for all views. Then we construct bipartite graph between raw data points and these salient points. These generated points play an important role in capturing the manifold of the original views. Then, the graph of all the views are combined together using a local manifold fusion method. Finally, we run spectral clustering on the resulting fused graph. There are several benefits of our method: **First**, manifold fusion preserves the manifold structure of all the views; **Second**, the construction of the bipartite graph is very efficient; **Third**, by exploring the special structure of the bipartite graph, spectral analysis on it is also very efficient; **Fourth**, our method also output cluster indicator of the salient points, which enables us to handle the out-of-sample problem efficiently. Additionally, we have conducted extensive experiments on five ‘ benchmark data sets, which demonstrate the effectiveness and efficiency of our proposed method comparing to the state-of-the-art methods.

The remainder of this chapter is organized as follows: we first introduce basic notations and concepts of spectral clustering in Section 2. In Section 3, details of our proposed large-scale multi-view spectral clustering method is presented. All the experimental results are shown in Section 4. Finally, we conclude our work in Section 5.

4.2 Background and Notations

In this section, we will briefly introduce the notations and the spectral clustering framework. Let $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$ denote the data matrix, where n is the number of data points and d is dimension of features. Each data point $x_i \in \mathbb{R}^d$ belongs to one of K classes $C = \{c_1, \ldots, c_K\}$. Given the whole dataset X, each data point is represented as a vertex on the affinity graph and each edge represents the affinity relation of one pair of vertexes. In practice, the k-NN graph are usually used.

Specifically, x_i and x_j are connected if at least one of them is among the k nearest neighbours of the other in the given measured (usually Euclidean distance). The weight of the edge between x_i and x_j is defined as:

$$w_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right), & \text{if } x_i \text{ and } x_j \text{ are connected} \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where σ is the bandwidth parameter. Note that we use Gaussian Kernel for example, this method is also applicable to other types of kernel. Thus, $W = \{w_{ij}\} \in \mathbb{R}^{n\times n}, \forall i, j \in 1, \ldots, n$ is the adjacent matrix of the graph and it is a symmetric undirected graph. Let $D \in \mathbb{R}^{n\times n}$ be the degree matrix whose i-th diagonal element is $d_{ii} = \sum_{j=1}^{n} w_{ij}$. Let L denote the normalized graph Laplacian matrix, then it is defined as:

$$L = I - D^{-1/2} W D^{-1/2} \tag{4.2}$$

The objective function of the normalized spectral clustering [53] is defined as:

$$\min_{G^T G = I} \operatorname{Tr}(G^T L G), \tag{4.3}$$

where $G \in \mathbb{R}^{n\times K}$ is the class indicator matrix of all data. The solution of G in Eq. (4.3) is the K smallest eigen vectors of L.

4.2.1 Multi-view Spectral Clustering Revisit

For multi-view data, let V be the number of views and $X^{(1)}, \ldots, X^{(V)}$ be the data matrix of each view, where $X^{(v)} \in \mathbb{R}^{n\times d^{(v)}}$ for $v \in 1 \ldots, V$ and $d^{(v)}$ is the feature dimension of the v-th view. Let $L^{(1)}, \ldots, L^{(V)} \in \mathbb{R}^{n\times n}$ denote the normalized Laplacian matrices of each view, respectively. Two important questions that are needed to be answered by multi-view approaches are how to reach consensus of the

results and how to express the relationship of all the views. There are several forms for the multi-view spectral clustering [65, 66]. We use the following form:

$$\min_{G^TG=I, a^{(v)}} J_1(G, a^{(v)}) = \sum_{v=1}^{V} (a^{(v)})^r \operatorname{Tr}(G^T L^{(v)} G),$$
$$\text{s.t. } \sum_{v=1}^{V} a^{(v)} = 1, a^{(v)} \geq 0, \tag{4.4}$$

where $a^{(v)}$ is the non-negative normalized weight factor for the v-th view and r is a scalar to control the distribution of different weights among different views. Here, we try to find a consensus result G among all the views. This unique consensus eliminates the need for computing the local results for each view and the computation cost of communicating back and forth between local results and the global result e.g. [65]. To further explain the inter-view relation, we rewrite Eq. (4.4) as:

$$\min_{G^TG=I, a^{(v)}} J_2(G, a^{(v)}) = \operatorname{Tr}(G^T L G),$$
$$\text{s.t. } \sum_{v=1}^{V} a^{(v)} = 1, a^{(v)} \geq 0, \tag{4.5}$$

where $L = \sum_{v=1}^{V} (a^{(v)})^r L^{(v)}$. Here, L can be regarded as local manifold fusion of all the views.

Equation (4.5) can be solve by iterative optimization techniques. However, to construct the graphs for all the views and to solve the equation is time consuming. The computational complexity is about $O(TKn^2 + \sum_{v=1} Vnd_v^2)$, where T is the number of iterations.

4.3 Methodology

In this section, we present an efficient approximation algorithm that can be applied to large-scale graph construction. Then, an efficient clustering algorithm

is proposed for large-scale multi-view spectral clustering. Finally, we extended our method to handle the out-of-sample problem.

4.3.1 Large-Scale Graph Construction

In order to reduce the computational cost of multi-view spectral clustering, we introduce a fast approximation algorithm. The idea is to use a small set of data points $U = [U_1, \ldots, U_m] \in \mathbb{R}^{m \times d}$ to capture the manifold structure, where each u_k is called a salient point. Then a bipartite graph is constructed between the raw data points and the salient points. By utilizing the structure of the bipartite graph, the graph construction and spectral analysis can be perform very efficient.

The salient points can be chosen by random sampling from raw data points or using lightweight clustering methods such as k-means. We find that the salient points generated by k-means have stronger representation power compared to sampling ones, where fewer points are needed for the same level of performance. However, in multi-view data, different views will generate different salient points if we run k-means independently on each view, which makes manifold fusion impossible. Therefore, we generate salient points on concatenated all the features and then separate resulting points into different views. This process can generate uniform salient points for different views, which will simplified the process of clustering.

With the generated points, the k-NN graph is constructed between the raw data and the salient points. We further constrain that connections are only allowed between raw data point and salient point. This constraint results in a bipartite graph between raw data X and salient points U. And the weight of each edge is defined as

$$Z_{ij} = \frac{K(x_i, u_j)}{\sum_{k \in \Phi_i} K(x_i, u_k)}, \forall j \in \Phi_i, \tag{4.6}$$

where $K()$ is a given kernel function (e.g. Gaussian Kernel in Eq. (4.1)), $\Phi_i \subset \{1 \dots m\}$ denotes the indexes of s nearest neighbours of x_i in U.

For the v-th view, the affinity matrix becomes

$$W^{(v)} = \begin{bmatrix} 0 & Z^{(v)} \\ {Z^{(v)}}^T & 0 \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}.$$

The degree matrix becomes

$$D^{(v)} = \begin{bmatrix} D_r^{(v)} & 0 \\ 0 & D_c^{(v)} \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)},$$

where D_r is a diagonal matrix of whose diagonal elements are row sums of Z and D_c is a diagonal matrix of whose diagonal elements are column sums of Z. Since Z is by definition row normalized, we have $D_r = I_n$, where I_n is the n by n identity matrix. The construction of the graph is extremely efficient since now we only need to consider $O(mn)$ distances. However, directly computing eigenvectors of L in Eq. (4.4) is still time consuming. Therefore, we need to transform the problem to utilize the structure of bipartite graph.

4.3.2 Multi-view Spectral Clustering Algorithm

By utilize the bipartite graph, we can obtain an algorithm that can optimize the cluster indicator of raw data points and salient points simultaneously. We name this algorithm **Multi-view Spectral Clustering (MVSC)**. We first propose our alternative optimization framework for solving Eq. (4.5). With all the $a^{(v)}$ are initialized to be equal, i.e. $a^{(v)} = 1/V$ for $v \in 1 \dots V$, we solve Eq. (4.5) in iterations of two following steps.

First, we fix $a^{(v)}$ and then solve G, where the objective function become:

$$\min_{G^T G = I} J_2(G) = \mathrm{Tr}(G^T L G), \tag{4.7}$$

which is equivalent to original spectral clustering. The solution of G is obtained by compute K smallest eigenvectors of L.

Second, we fix G and then solve $a^{(v)}$. Let $h^{(v)} = \mathrm{Tr}(G^T L^{(v)} G)$, then the Eq. (4.4) can be rewritten as:

$$\min_{a^{(v)}} \sum_{v=1}^{V} (a^{(v)})^r h^{(v)}, \text{ s.t. } \sum_{v=1}^{V} a^{(v)} = 1, a^{(v)} \geq 0, \tag{4.8}$$

Thus, using method of Lagrange multiplier, Eq. (4.8) becomes:

$$\min_{a^{(v)}} \sum_{v=1}^{V} (a^{(v)})^r h^{(v)} - \beta(\sum_{v=1}^{V} a^{(v)} - 1), \tag{4.9}$$

where β is the Lagrange multiplier. With simple algebraic manipulations, we get

$$a^{(v)} = \frac{\left(r(h^{(v)}\right)^{\frac{1}{1-r}}}{\sum_{v=1}^{V} (rh^{(v)})^{\frac{1}{1-r}}}. \tag{4.10}$$

The first sub-problem (Eq. (4.7)) tries to minimize $J_2(G) = \mathrm{Tr}(G^T L G)$, which takes $O(cn^2)$ for general case. Fortunately, we can reduce the complexity by using the following theorem.

Theorem 5. *: Solving $J_2(G) = \mathrm{Tr}(G^T L G)$ is equivalent to compute the singular vectors of Z corresponding to K largest singular values.*

Proof. Let $S^{(v)} = (D^{(v)})^{-1/2} W^{(v)} (D^{(v)})^{-1/2}$. The objective function $J_2(G)$ can be rewritten as

$$\begin{aligned} J_2(G) &= \mathrm{Tr}(G^T L G) \\ &= \mathrm{Tr}(G^T \sum_{v=1}^{V} (a^{(v)})^r L^{(v)} G) \\ &= \mathrm{Tr}(G^T \left(\sum_{v=1}^{V} (a^{(v)})^r (I - S^{(v)}) \right) G) \\ &= \mathrm{Tr}(\sum_{v=1}^{V} (a^{(v)})^r G^T G \\ &\qquad - G^T \left(\sum_{v=1}^{V} (a^{(v)})^r S^{(v)} \right) G) \\ &= n \sum_{v=1}^{V} (a^{(v)})^r - \mathrm{Tr}(G^T S G), \end{aligned} \tag{4.11}$$

where $S = \sum_{v=1}^{V} (a^{(v)})^r S^{(v)}$. Then, minimizing J_2 with respect to G is equivalent to the following equation

$$\max_{G^T G = I} \mathrm{Tr}(G^T S G) \tag{4.12}$$

The solution of G is the eigenvectors corresponding to the K largest eigenvalues. We can use the structure of S to transform the problem of computing the eigenvectors of S to that of computing the eigenvectors of Z. Let $G = [G_X^T, G_U^T]^T$, where G_X, G_U

are rows corresponding to raw data and salient points respectively. Therefore, the objective function in Eq. (4.12) becomes

$$\begin{aligned}\mathrm{Tr}(G^T S G) &= \mathrm{Tr}\left(\begin{bmatrix} G_X \\ G_U \end{bmatrix}^T S \begin{bmatrix} G_X \\ G_U \end{bmatrix}\right) \\ &= \mathrm{Tr}\left(\begin{bmatrix} G_X \\ G_U \end{bmatrix}^T \begin{bmatrix} 0 & \hat{Z} \\ (\hat{Z})^T & 0 \end{bmatrix} \begin{bmatrix} G_X \\ G_U \end{bmatrix}\right) \\ &= \mathrm{Tr}\left(2 G_X^T \hat{Z} G_U\right), \end{aligned} \tag{4.13}$$

where $\hat{Z} = \sum_{v=1}^{V} (a^{(v)})^r \hat{Z}^{(v)}$ and $\hat{Z}^{(v)} = (D_r^{(v)})^{-\frac{1}{2}} Z^{(v)} (D_c^{(v)})^{-\frac{1}{2}} = Z^{(v)} (D_c^{(v)})^{-\frac{1}{2}}$. Thus, solving Eq. (4.12) is equivalent to computing the left and right singular vectors corresponding to the K largest singular values of $\hat{Z}$

$$svd(\hat{Z}) = G_X \Sigma G_U^T, \tag{4.14}$$

where $svd()$ is the Singular Value Decomposition (SVD) operator, $\Sigma = diag(\sigma_1, \ldots, \sigma_K)$ and $\sigma_1 \geq \sigma_2, \ldots, \sigma_K \geq 0$ are the singular values of $\hat{Z}$. □

With Theorem 5, we can solve the whole problem very efficiently. The whole algorithm is summarized in Alg. 5.

Computational analysis. The proposed Multi-view Spectral Clustering - (MVSC) consists of three stages: 1) generating salient points using k-means, 2) constructing graph Z and 3) optimization by iteratively solving the clustering problem. The first stage takes $O(t_1 nmd)$ time, where t_1 is the number of iterations for running k-means and $d = \sum_{v=1}^{V} d^{(v)}$. The second stage takes $O(nmd)$ to construct the graph Z, while constructing a normal k-NN graph of n vertexes takes $O(n^2 d)$. The third stage takes $O(t_2 nm^2)$, where t_2 is the number of iterations. Note that the optimization stage is much faster than clustering on a normal n by n graph, which takes

$O(Kn^2)$ time. So the overall time complexity is approximately $O(t_1nmd + t_2nm^2)$. Since $m, d \ll n$, this is nearly linear to n. The computational cost is summarized in Table 4.1.

Table 4.1: Summary of computational complexity.

Stages	1 and 2	3	Total
Normal graph	$O(n^2d)$	$O(Kn^2)$	$O(n^2d + Kn^2)$
Bipartite graph	$O(t_1nmd)$	$O(t_2nm^2)$	$O(t_1nmd + t_2nm^2)$

Convergence analysis. The original problem Eq. (4.4) is not a joint convex problem of $a^{(v)}$ and G. Hence, there is no guarantee for obtaining a global solution. Since we divide the original problem into two sub-problems and each of them is convex problem. The proposed method will converge to a local solution. In all our experiments, the process always converges in less than 10 iterations.

Parameter r. Another advantage of our approach is using the parameter r, which controls the fusion weights of all the views by only one parameter. Some previous methods just simply assume equal weights [65] or tuning one parameter for each view [73]. The effect of r ranges from assigning equal weights to all views when $r = \infty$ to assigning all the weights to one best view when $r = 1$. By tuning r between $(1, \infty)$ we can reach a balance between all the views.

Algorithm 5 Multi-view Spectral Clustering (MVSC)

Input: Data matrix of all views $X^{(v)} \in \mathbb{R}^{n \times d^{(v)}}$ for $v \in 1 \dots V$, Number of classes K, Number of salient points m, parameter r.

Output: Cluster labels Y of each data points, all salient points U and cluster labels of all salient points.

Generate m salient points using k-means on concatenate features;

Compute affinity matrix $Z^{(v)}$ of each view.

Compute Laplacian $L^{(v)}$ of each view;

Initialize $a^{(v)} = 1/K$;

repeat

    Compute G by using Eq. (4.14);

    Update $a^{(v)}$ by using Eq. (4.8);

until Converges.

Treat each row of G as new representation of each data point and compute the clustering labels Y by using k-means algorithm.

4.3.3 Out-of-sample Problem

In general, spectral clustering methods only work on the training data. Most methods do not provide clear extension to deal with out-of-sample points (a.k.a. test data). In contrast, our method can be easily extended to handle test data. Recall that when carrying out clustering on training data, we also get the feature vectors and clustering labels for the salient points. Therefore, we simple find the k nearest neighbours of test data among salient points and propagate the labels to the test data. The k-NN algorithm can be done in $O(md)$ computational cost for each data

point. Hence, p test data points can be clustered in $O(pmd)$ computational cost. This computational cost is far lower than carried out k-NN on the training data ($O(pnd)$).

4.4 Experiment

In this section, we conduct several experiments to evaluate the performance of the proposed methods on five benchmarks datasets. These datasets are summarized in Table 4.2. All our experiments are conducted on a desktop computer with a 3.4GHz Intel Core i7 CPU and 12GB RAM, MatLab 2012a (64bit).

4.4.1 Data Set Description

Handwritten (HW)[1] is a dataset of handwritten digits of 0 to 9 from UCI machine learning repository [74]. It consists of 2000 data points. We use all the 6 published features including 76 Fourier coefficients of the character shapes (FOU), 216 profile correlations (FAC), 64 Karhunen-love coefficients (KAR), 240 pixel averages in 2×3 windows (Pix), 47 Zernike moment (ZER) and 6 morphological (MOR) features.

Caltech-101 [75] image data set consists of 101 categories of images for object recognition problem. We follow previous work [76] and select the widely used 7 classes, i.e. Face, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign and Windsor-Chair and get 1474 images, which we called **Caltech7 (Cal7)**. We also select a larger set named **Caltech20 (Cal20)** which contains totally 2386 images of 20 classes: Face, Leopards, Motorbikes, Binocular, Brain, Camera, Car-Side, Dolla-Bill, Ferry, Garfield, Hedgehog, Pagoda, Rhino, Snoopy, Stapler, Stop-Sign, Water-Lilly, Windsor-Chair, Wrench and Yin-yang. Five features are extracted from all the images: i.e. 48 dimension Gabor feature, 40 dimension wavelet moments (WM), 254 dimension CENTRIST

[1] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

feature, 1984 dimension HOG feature, 512 dimension GIST feature, and 928 dimension LBP feature.

Reuters[2] consists of documents that are written in five different languages and their translations. All the documents are categorized in to 6 classes. We use the subset that are written in English all their translations in all the other 4 languages (French, German, Spanish and Italian).

NUS-WIDE-Object (NUS) [77] is a dataset for object recognition which consists of 30000 images in 31 classes. We use 5 features provided by the website [3], i.e. 65 dimension color Histogram (CH), 226 dimension color moments (CM), 145 dimension color correlation (CORR), 74 dimension edge distribution and 129 wavelet texture.

Animal with attributes (**AWA**)[4] is a data set of animal images. It consists of 50 kinds of animals described in 6 features. We randomly sample 80 images for each class and get 4000 images in total. All the published features are used: Color Histogram (CQ, dim 2688), Local Self-Similarity (LSS, dim 2000), Pyramid HOG (PHOG, dim 252), SIFT (dim 2000), Color SIFT (RGSIFT, dim 2000) and SURF (dim 2000).

4.5 Clutering Evaluation

In this subsection, we first evaluate the capability of the proposed multi-view clustering method on 5 datasets: HW, Caltech7, Caltech20, Reuters and NUS. We compare the proposed methods with three other state-of-art approaches as stated bellow:

[2]https://archive.ics.uci.edu/ml/datasets.html

[3]http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[4]http://attributes.kyb.tuebingen.mpg.de/

Table 4.2: Summary of the multi-view datasets used in our experiments.

No.	HW	Caltech7/20	Reuters	NUS	AWA
1	Pix(240)	Gabor(48)	English(21531)	CH(65)	CQ(2688)
2	Fou(76)	WM(40)	France(24892)	CM(226)	LSS(2000)
3	Fac(216)	CENTRIST(254)	German(34251)	CORR(145)	PHOG(252)
4	ZER(47)	HOG(1984)	Italian(15506)	EDH(74)	SIFT(2000)
5	KAR(64)	GIST(512)	Spanish(11547)	WT(129)	RGSIFT(2000)
6	MOR(6)	LBP(928)	-	-	SURF(2000)
#data	2000	1474/2386	18758	26315	4000
#classes	20	7/20	6	31	50

Single view Spectral Clustering (**SC**): Running spectral clustering on each single view [53].

Feature Concatenation Spectral Clustering (**ConSC**): Concatenating features of all the views and run spectral clustering on the resulted feature [65].

Co-regularized Spectral Clustering (**CoregSC**): one of the state-of-the-art multi-view spectral clustering method proposed in [65].

Multi-Modal Spectral Clustering (**MMSC**): another recent multi-view clustering method proposed in [66].

Multi-view Spectral Clustering (**MVSC**): this is the proposed method in Alg. (5).

For fair comparison, we download the source code from the authors' website and follow their experimental setting and the parameter tuning steps in their chapter. And we use Gaussian kernel for all the experiments except for the Reuters dataset, where we use linear kernel. We search the parameter r in logarithm form ($\log_{10} r$ from 0.1 to 2 with step size 0.2. We also set $m = 400$ and construct 8-nearest-neighbour graph between raw All the experiments are repeated for 10 times and average results are reported. For the experimental results, we report three metrics [78]: **mean purity**, **mean mutual information (NMI)** and **mean running time**.

Table 4.3: Clustering purity comparison on all data sets. "OM" means "Out-of-memory error" while running the experiment.

Data set	HW	Cal7	Cal20	Reuters	NUS
SC(1)	75.12%	79.22%	68.13%	53.10%	15.98%
SC(2)	75.44%	79.85%	68.13%	54.86%	16.13%
SC(3)	76.39%	79.36%	69.09%	56.92%	15.78%
SC(4)	73.47%	80.56%	67.01%	53.82%	16.29%
SC(5)	75.84%	80.48%	67.99%	56.79%	16.44%
SC(6)	78.89%	79.97%	66.90%	-	-
ConcatSC	59.33%	77.96%	60.33%	**56.70%**	26.81%
CoRegSC	**82.23%**	**83.71%**	**76.11%**	55.23%	**26.49%**
MMSC	75.84%	84.47%	69.04%	39.01%	OM
Proposed	**84.41%**	**84.66%**	**74.06%**	**57.73%**	**28.21%**

Table 4.4: Clustering NMI comparison on all data sets. "OM" means "Out-of-memory error" while running the experiment.

Data set	HW	Cal7	Cal20	Reuters	NUS
SC(1)	0.7589	0.4189	0.4842	0.3099	0.0398
SC(2)	0.7549	0.4239	0.4813	0.3033	0.0419
SC(3)	0.7556	0.4217	0.4848	0.3039	0.0403
SC(4)	0.7547	0.4220	0.4816	0.3123	0.0432
SC(5)	0.7576	0.4206	0.4830	0.3078	0.0429
SC(6)	0.7577	0.4190	0.4830	-	-
ConcatSC	0.5795	0.2734	0.3590	0.3228	0.1421
CoRegSC	**0.8358**	0.5253	**0.6107**	**0.3261**	**0.1428**
MMSC	0.7920	**0.5638**	**0.5938**	0.1335	OM
Proposed	**0.8324**	**0.5586**	0.5698	**0.3567**	**0.1493**

Table 4.3 and Table 4.4 show clustering purity and NMI respectively, while Table 4.5 shows the running time of all the methods. In general, the multi-view methods can achieve better results than the single view algorithms. Additionally, our proposed method MVSC constantly outperforms the single view methods and achieves comparable or even better results than the other multi-view methods. For running time comparison in Table 4.5, the proposed method is up to several orders of magnitude faster than the baseline methods. The gap is even larger in the large datasets. The other benefits of the proposed method is low space complexity. In fact,

Table 4.5: Running time comparison on all data sets (seconds). "OM" means "Out-of-memory error" while running the experiment.

Data set	HW	Cal7	Cal20	Reuters	NUS
SC(1)	1.74	10.94	29.18	556.98	852.07
SC(2)	1.54	10.30	29.27	443.92	580.36
SC(3)	1.53	10.20	29.32	422.91	478.88
SC(4)	1.53	10.16	29.15	354.68	527.62
SC(5)	1.58	10.32	29.32	307.62	633.01
SC(6)	1.53	10.24	29.33	-	-
ConcatSC	2.20	11.00	26.19	556.73	2172.90
CoRegSC	16.42	61.78	180.65	7074.17	56327.26
MMSC	6.13	27.62	80.25	14556.13	OM
Proposed	**0.84**	**1.21**	**2.26**	**135.48**	**19.34**

nearly all the baseline methods raise out-of-memory exception when number of data points are more than 40,000 while the proposed method can easily handle more than 100,000 samples at once.

4.6 Out-of-sample Problem

In this subsection, we consider the out-of-sample problem. Experiments are conducted on AWA dataset. Five fold cross-validation is used and we report the mean purity, the mean NMI and the mean testing time. At each fold, 1/5 of the data are used as in-sample clustering like that in the previous subsection and the other 4/5 are used for the out-of-sample test. For the out-of-sample test, the data and the estimated cluster labels of the in-sample clustering are used as training data for the model. Here we compare two situations: **1)** training model with the whole raw in-sample data; **2)** training model with the generated salient points. Two kinds of models are trained in both situation: **Linear Regression (LR)** and **Nearest Neighbour (1NN)**. The corresponding models trained on the salient points are called **Salient 1 Nearest Neighbour (Sa1NN)** and **Salient Linear Regression (SaLR)** respectively. We compare the proposed method with a third baseline method

Spectral Embedded Clustering (SEC) [67]. Since the data have several views, we train and apply models on each view and use simple voting scheme to decide the final cluster label for each testing sample.

Table 4.6: Results of out-of-sample test om AWA.

Method	1NN	LR	SEC	Sa1NN	SaLR
Purity	8.13%	7.22%	7.79%	8.37%	7.31%
NMI	0.1395	0.1124	0.1252	0.1490	0.1120
Time (s)	972.53	0.97	0.99	436.45	0.94

Table 4.6 shows the testing performance of all the methods. The first two rows of the table are purity and NMI, respectively, while the third row shows the testing time. We can observe that the purity of the salient-point-based models are comparable or even better than the raw-data-based models. The testing time of Sa1NN is much less than the 1NN model. This is reasonable since the computational complexity of 1NN algorithm is proportional to the number of training samples. All these results demonstrate that we can achieve comparable performance using models trained only on the salient points.

4.7 Conclusion

In this chapter, we propose a novel large-scale multi-view spectral clustering method based on bipartite graph, named MVSC. Given a multi-view data set with n data points, MVSC select m uniform salient points among all the views to represent the manifold structures of all the features. For each view, one sub-bipartite-graph is constructed between the raw data points and the generated salient points. We use local manifold fusion to generate a fused bipartite graph to integrate information of all the sub-graph. By exploring the structure of the bipartite graph, the clustering

process can be accelerated significantly. The computational complexity is close to linear to the number of data points. For the clustering results, we not only obtain cluster labels for the training data but also cluster labels for the salient points. The later information has been used to handle the out-of-sample problem in low computational cost. Extensive experiments on five benchmark data sets demonstrate that our proposed method is up to several orders of magnitude faster than the state-of-the-art methods, while preserving the comparable or even better accuracy.

CHAPTER 5

INSTRUMENT TRACKING VIA ONLINE LEARNING IN RETINAL MICROSURGERY

In this chapter, we turn to the problem of real-time processing of sequential data. We investigate the problem of visual tracking of instruments in microsurgery, which is a typical example of image sequence task. An algorithm is proposed to achieve better accuracy than the state-of-the-art approaches while running in video frame rate [79].

5.1 Introduction

Retinal microsurgery (RM) is an important treatment for sight-threatening conditions. The procedure is performed by a surgeon using a microscope for visualization and manipulating a set of surgical instruments. The operating surgeon faces several difficulties such as indirect visualization of the surgical target, hand tremors and lack of tactile feedback. To overcome these difficulties, new techniques have been developed. Accurate visual tracking of surgical tools in microscopic images is an important technique to complement the previously developed smart tools. In this chapter, we focus on the task of robust visual tracking of instruments in in-vivo RM monocular image sequences.

This task is challenging due to the great variability in the appearance of surgical tools because of illumination and other factors. Many existing methods focus on training the appearance model based on color features or the instrument geometry [80, 81, 82, 83]. However, these methods often perform poorly under complex appearance

changes due to their oversimplified appearance models. Sznitman et al. proposed an approach, namely Data-Driven Visual Tracking (DDVT) [84], which integrates an instrument detector based on deformable features with a simple gradient-based tracker. DDVT is able to run in video frame rate and achieves state-of-the-art results on challenging human in-vivo surgery datasets. To our best knowledge, DDVT is by far the best visual tracking approach in RM. However, there are two drawbacks to DDVT. First, it needs manually labelled instrument positions in many video frames for training the offline detector. Second, it performs poorly in handling appearance changes that were not observed in the training sequences and could not be modelled by the trained offline detector.

Currently, it draws more and more attentions to integrate online learning techniques in visual tracking system [85, 86]. How to extract new reliable samples without corrupting the current model is a key problem to this kind of systems. Therefore, many techniques have been exploited to constrain the learning process [87, 88]. However, many existing models are not robust enough to apply on RM tracking problem due to the challenges discussed above.

To this end, we propose a new approach based on online learning—Instrument Tracker via Online Learning (ITOL). In this approach, we adopt the paradigm of combining tracking and detection in the same framework [89, 90]. ITOL uses a robust gradient-based tracker capable of failure detection as the basic tracker. Then, a cascade appearance classifier is used as the instrument detector. The appearance model of the detector is initialized by manually clicking the instrument position in the first frame. It is adaptively trained and updated on the fly. Samples for online updating are collected by a filtering process, which selects "unfamiliar" positive samples and "hard" negative samples. The obtained training set is used to augment the model of the detector and prevent the detector from making the similar mis-

takes. The performance of the proposed approach is evaluated in three human in-vivo retinal microsurgery videos and one laparoscopy image sequence. The experimental results demonstrate that our method significantly outperforms the state-of-the-art approaches.

The rest of this chapter is organized as follows: Section 5.2 introduces the framework and each components of our approach. Then we present our experimental results in Section 5.3 and conclude the proposed approach in Section 5.4.

5.2 Method

In this section, we will detail our proposed method ITOL. Methods for visual tracking usually fall into two groups: tracking through local optimization and tracking by detection [81]. Tracking through local optimization is fast, accurate and able to handle appearance changes of the target. However, continuous template updating is needed in order to maintain accurate position tracking when there are significant changes in target appearance [84]. Tracking by detection has the advantage of being able to handle target disappearance, but the ability of detection is limited by the training data.

Instrument tracking is challenging due to often unexpected appearance changes and extreme deformations of the instrument. We use a multi-component tracking framework to address these problems. A flowchart diagram of the framework is shown in Fig. 5.1. First, a robust gradient-based **tracker** with the ability of failure detection is used to handle unexpected appearance changes. Then an instrument **detector** is adopted to compensate for tracking loss and it automatically re-initializes the tracker when the instrument reappears after disappearance or tracking loss. To provide more reliable tracking results, outputs of the tracker and the detector will be integrated into a unique target position by a component named **integrator**. Finally, a component

named **sample expert** will be used to efficiently select image patches for online updating of appearance model of the detector. In the whole framework, we only need to manually click the position of the instrument in the first frame for training data. Then, the tracking system is fully automatic. Details of each component of the system will be discussed in the following sections.

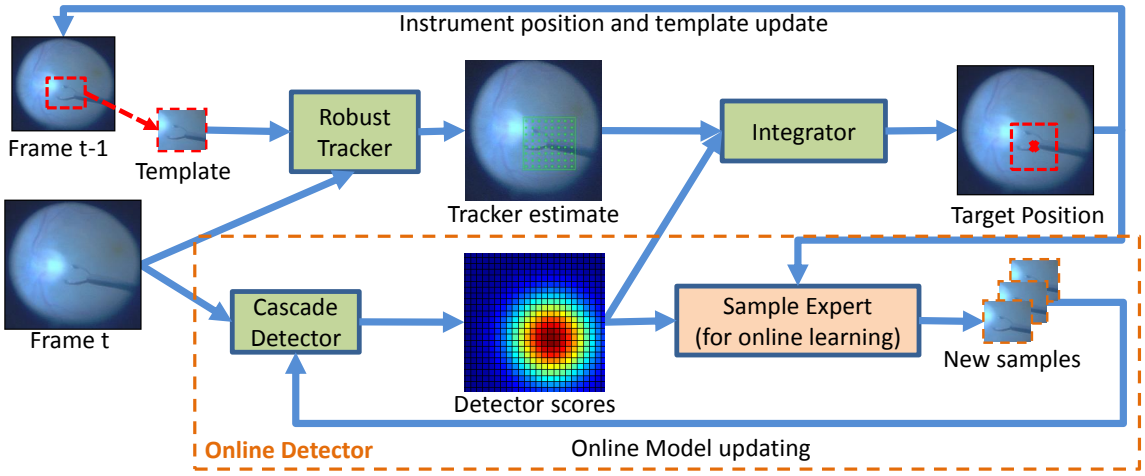


Figure 5.1: Diagram of our ITOL framework.

5.2.1 Robust Tracker

The tracker is used to handle instrument appearance changes and bring in new appearance samples. In many cases, although the appearance in the current frame is new to the current model of the detector, it is gradually adapted over time from seen samples. Since we use a gradient based tracker, which is only concerned with similarity between two consecutive frames, it can adaptively collect new appearance samples while tracking. The tracker is based on the Median Flow (MF) algorithm [91]. In the Median Flow tracker, the target is represented by a bounding box around it. For robustness, the bounding box is divided into a $k \times k$ grid ($k = 10$ in our

experiments), where each cell of the grid is tracked by the pyramidal L-K algorithm [92]. The displacement of the target is voted by 50% of the most reliable cells. The reliability level of a cell is measured by normalized cross-correlation (NCC).MF also uses a quantity named Forward-Backward (FB) error for failure detection. The tracking is performed both forward and backward along the time axis and the FB error is computed based on the discrepancies between these two trajectories of the target [91]. Since the instrument sometimes move severely or is out of view, this failure detection ability is critical to prevent the tracker from importing false samples.

5.2.2 Cascade Detector

The gradient-based MF tracker assumes that the target is always in view and under continuous changes. In practice, instruments or tools during RM often undergo large appearance changes, which breaks the assumption. An online detector is developed to compensate for this shortcoming of the tracker and to re-initialize the tracker when an instrument reappears after loss. The detector scans the current frame by sliding window and decides whether the target is present in each window. A complex object detector often requires high computational cost, which makes it impossible for real-time surgical tracking. This problem is addressed by combining successively more complex classifiers in a cascade structure, which rejects most negative windows in the early stages of the cascade thus increasing the processing speed of the detector [93].

In our method, each frame is scanned by the detector at multiple scales using sliding window. All the candidate bounding boxes will be resized to the same size. Inspired by [90], we use a three-stage detector. The first stage is a variance filter that checks if the variance of the patch is under certain threshold related to the variance of trained positive samples. The variance filter can be evaluated efficiently by using

integral images [93]. The second stage is random ferns (forest) [94] on patches for comparing the pixel values. Pixels in a patch are first divided into several groups. The probability is then computed for each group based on the number of times that the same feature combination appeared in previous frames as positive or negative examples. The final confidence score is computed by averaging the probabilities of each group. The third stage is a 1-Nearest-Neighbour (1NN) classifier using Normalized Correlation Coefficient (NCC) as the distance between the candidate patch and two sets of patches: positive patches and negative patches. Usually, the first two stages are able to reject more than 95% of the candidate windows, which makes the detector very efficient. In fact, this detector is able to run at nearly 30fps in our experiments.

5.2.3 Integrator

As discussed above, the detector and the tracker have their respective advantages and disadvantages. Therefore, we use the integrator to integrates their outputs to achieve an optimal estimation. The rules for this integration are: 1) If neither the tracker nor the detector output any positions, the target is declared as not visible; 2) Otherwise, all the outputs of the tracker and the detector are clustered into one by their scores. Suppose s_+ is the similarity between a candidate patch and its nearest neighbour in the positive sample set and s_- is the similarity between the patch and its nearest neighbour in the negative sample set, and $\rho = \frac{s_+}{s_-}$. Then the score of the patch is defined as $s = \frac{1}{1+\rho}$.

5.2.4 Online Updating of Detector's Model

The sample expert is designed to select new training samples for online model updating of the detector. Online updating make the detector capable of handling unexpected appearance changes and more robust to the noises. Given new samples,

the updating process is straightforward. For random ferns, the probability of each branch is updated by adding the results of the pixel comparison. The 1NN classifier simply adds new samples to its sample sets.

The online learning method is detailed in the following. To prevent false positive samples, the sample expert use higher threshold than the detector. Then we consider these bounding boxes as potential positive samples. Starting from the output of the integrator, the sample expert will generate the new positive samples by choosing bounding boxes that are very close to the output one. Second, we filter them by our 1NN classifier and only accept the samples that are rejected by the 1NN classifier. The second step has two effects: 1) It rules out those "easy" samples to avoid redundancy; 2) The remaining samples are "new" enough so that the model will improve very rapidly. In order to accelerate the growth of the model, positive sample are rotated and blurred to generate more data. For negative examples, a common practice is focusing on "hard" samples. Therefore, only samples that have passed the first two stages of the detector and far away from the output are considered candidates of negative samples.

5.3 Experiment and Results

In this section, we conduct experiments to evaluate ITOL on two public datasets: **Retina Microsurgery Dataset** and **Laparoscopy Sequence** [84].

- **Retina Microsurgey Dataset** consists of 3 sequences of in-vivo vitreoretinal surgery, which contains a total of 1171 images (640×480 pixels). See Fig. 5.2 for examples. These sequences are challenging due to variations in illumination type and quantity, light source position and the presence of blur and shadows.

- **Laparoscopy Sequence** consists of 1000 images with labelled locations of the tool tip. The original video is from Youtube. There are two instruments in each image, hence there are roughly 2000 instrument locations.

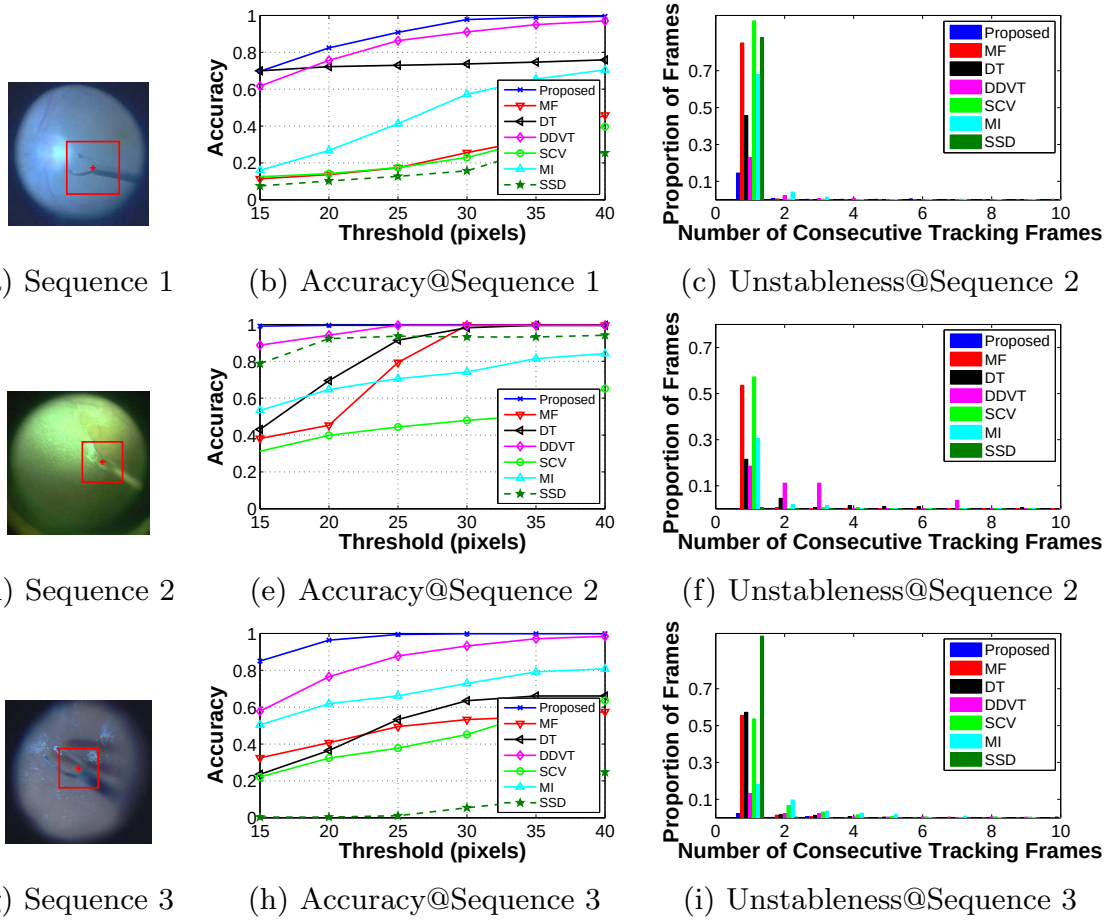
We compare our method **ITOL** with four baseline methods: **DDVT** [84], **SCV** [95], **MI** [83], **SSD** [96]. We also compare two components used in the proposed method: **Median-Flow (MF)** and **Detector-Tracker (DT)**. **MF** is the gradient-based tracker that we used. **DT** is MF plus the cascade detector without online model updating. For fair comparison, two measures are used by following the experimental setting of [84]: the accuracy on the thresholding distance to groundtruth and the number of the consecutive tracking frame. The accuracy is defined as the percentage of the detection within δ pixels of the groundtruth annotation. We vary δ from 15 to 40 in experiments (same as the setting in DDVT [84]).

The proposed method is implemented in Matlab. All experiments are conducted on a Desktop PC, 3.4GHz Intel Core i7-3770 and 12GB RAM. Our method runs at nearly 20fps and should run even faster implemented on parallel architecture (e.g. GPU or Mutlti-core).

5.3.1 Retina Microsurgery Dataset

The experimental results on the RM dataset are shown in Fig. 5.2. Results of each video sequence are shown in one row. In all the results, DDVT [84] outperforms the others except the proposed ITOL. ITOL also outperforms MF and DT, which validates the benefits of the online detector. Similar trends have been witness in all three videos where ITOL always achieves the best accuracy and unstableness. We accredit the advantages of the proposed ITOL to the online learning component that effectively updates the detector and makes it adapt to the appearance changes of instruments. One thing that is worth to note is DDVT uses the offline detector and

therefore requires sufficient amount of training data before tracking (e.g. 500 manually labelled frames [84]), while our method bases on online learning techniques and only requires one labelled position in the first frame as training data before tracking.



(a) Sequence 1 (b) Accuracy@Sequence 1 (c) Unstableness@Sequence 2

(d) Sequence 2 (e) Accuracy@Sequence 2 (f) Unstableness@Sequence 2

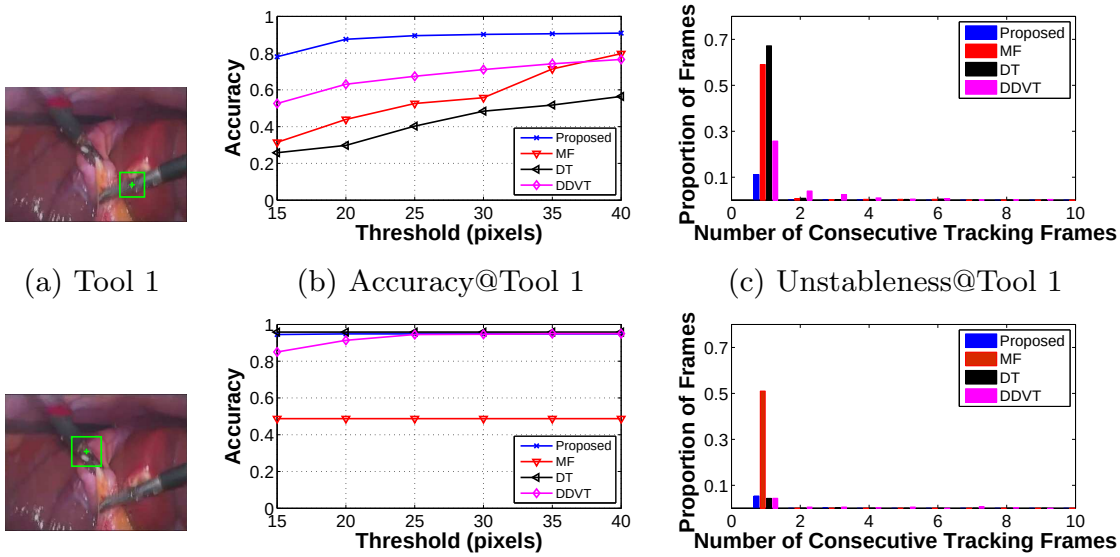
(g) Sequence 3 (h) Accuracy@Sequence 3 (i) Unstableness@Sequence 3

Figure 5.2: The results on Retina Microsurgery Dataset. For values of accuracy (the 2nd column), the higher the better. For values of unstableness (the 3rd column), the lower the better.

5.3.2 Laparoscopy Sequence

Finally, we also evaluate our method on the laparoscopic instrument sequence. The sequence is provided by [84]. DDTV uses the first 500 images for training and

the last 500 images for testing. For fair comparison, we follow the setting of [84] and use the last 500 images for testing. However, we only need one image frame for training before tracking because of the online learning technique. There are two tools in this video. For better visualization, we separately present the experimental results of two instruments in Fig. 5.3, one in each row. In the sequence, the first tool is under big changes in terms of the instrument structure and movement. Our method significantly outperforms DDVT [84] and two component methods. The second tool is relatively stable in shapes and positions in the whole testing image sequence. The results of the proposed approach are similar to those of the DDVT.



(a) Tool 1 (b) Accuracy@Tool 1 (c) Unstableness@Tool 1

(d)Tool 2 (e) Accuracy@Tool 2 (f) Unstableness@Tool 2

Figure 5.3: The results on Laparoscopy Sequence. For values of accuracy (the 2nd column), the higher the better. For values of unstableness (the 3rd column), the lower the better.

5.4 Conlcusion and Discussion

We proposed a novel approach, dubbed ITOL, for visual tracking of retinal instruments during in-vivo retinal microsurgery. Our method consists of four components: a robust gradient-based tracker, a cascade detector, an integrator and a sample expert. While the first three components make a robust and automatic tracker, the sample expert works to achieve online updating of the appearance model of the detector. ITOL only needs manually labelled position in the first frame and all remaining steps are fully automated, which makes it an approach needing much less user input than other existing methods. ITOL can also automatically re-initialize the tracker after failure. Experimental results on two video datasets demonstrate that the proposed method outperforms the state-of-the-art approaches. Our method makes tracking in RM much more feasible than before.

CHAPTER 6

Conclusions

This thesis aims at developing scalable machine learning and computer vision techniques for large-scale data. We investigate several typical type of data in the big data era including 1) high-dimensional data; 2) large-scale high-dimensional data; 3) large-scale multi-view data; 4) sequence data.

We have demonstrated, both in theory and practice, effective and efficient solutions with clear performance gains in extensive experiments on large-scale data. Specifically, we have developed the following methods:

Sub-representation for high-dimensional image data representation: The prior sparse or collaborative representations perform poor in high-dimensional data because of their computational cost. We have presented sub-representation, which used only a subset of data to estimate the representation coefficients as well as other unknowns. Our method can be combined with many variations of sparse/collaborative representation to handle problem like misalignment, occlusion. Our experiments have shown that sub-representation achieve same level of accuracy in various kinds of tasks like motion estimation, face recognition, with only a fraction of computational cost compared to the traditional sparse/collaborative representation.

Sub-selective Quantitation for nearest search of large-scale high dimensional data: We have developed a practical method for training unsupervised hashing function. For this problem, we have observed that training of many prior hashing algorithms mainly based on matrix computation on huge matrices. Then, we have developed the sub-selective quantitation approach, which use only a subset

of data to estimate the hash functions. Experimental results demonstrated that our approach can achieve up to dozens times of acceleration on running time compared to the state-of-the-art hashing approach.

Large-scale multi-view spectral clustering: We have addressed the problem of unsupervised learning and developed an efficient algorithm for spectral clustering on large-scale multi-view data. A novel graph construction approach has been proposed to efficiently approximate the original similarity graph on multi-view data with lower computational complexity. The key idea is the low-rank approximation of the fused graph of multi-view data. This approach not only accelerates the graph construction step but also the clustering step. In the later one, the most time consuming part is the singular value decomposition. We have conducted extensive experiments on several large data sets. The results demonstrated that we have achieve up to 1000 times of acceleration compared to the state-of-the-art multi-view clustering approaches.

Online learning for instrument tracking in microsurgery: In this part of thesis, we have proposed an online learning approach for visual tracking of instrument in microsurgery. This is a very important application in robot-assisted surgery and also a typical application on processing sequential data in real-time setting. Our experimental results showed that the proposed method have achieved better accuracy compared to the state-of-the-art approach while also running in real-time manner.

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[2] R. Rigamonti, M. Brown, V. Lepetit *et al.*, "Are sparse representations really relevant for image classification?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1545–1552.

[3] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 471–478.

[4] M. Yang, L. Zhang, and D. Zhang, "Efficient misalignment-robust representation for real-time face recognition," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 850–863.

[5] M. Yang, L. Zhang, D. Zhang, and S. Wang, "Relaxed collaborative representation for pattern classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2224–2231.

[6] Y. Li, C. Chen, and J. Huang, "Transformation-invariant collaborative sub-representation," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 3738–3743.

[7] C. Chen, Y. Li, W. Liu, and J. Huang, "Sirf: Simultaneous image registration and fusion in a unified framework," *arXiv preprint arXiv:1411.5065*, 2014.

[8] C. Chen, Y. Li, L. Axel, and J. Huang, "Real time dynamic mri with dynamic total variation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 138–145.

[9] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2760–2765.

[10] J. Huang, X. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[11] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 2, pp. 372–386, 2012.

[12] Y. Li, C. Chen, W. Liu, and J. Huang, "Subselective quantization for large-scale image search," in *AAAI*, 2014.

[13] B. R. Laura Balzano and R. Nowak, "High-dimensional matched subspace detection when data are missing," in *IEEE International Symposium on Information Theory Proceedings*, 2010.

[14] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[15] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[16] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Computer VisionECCV'92*. Springer, 1992, pp. 237–252.

[17] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[18] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[19] A. Martınez and R. Benavente, "The ar face database," *Rapport technique*, vol. 24, 1998.

[20] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[21] J. Huang, X. Huang, and D. Metaxas, "Learning with dynamic group sparsity," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 64–71.

[22] J. Huang, S. Zhang, H. Li, and D. Metaxas, "Composite splitting algorithms for convex optimization," *Computer Vision and Image Understanding*, vol. 115, no. 12, pp. 1610–1622, 2011.

[23] J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Medical Image Analysis*, vol. 15, no. 5, pp. 670–679, 2011.

[24] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Proc. CVPR*, 2008.

[25] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[26] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2143–2157, 2009.

[27] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for large-scale search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2393–2406, 2012.

[28] S. Korman and S. Avidan, "Coherency sensitive hashing," in *Proc. ICCV*, 2011.

[29] C. Strecha, A. M. Bronstein, M. M. Bronstein, and P. Fua, "Ldahash: Improved matching with smaller descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 66–78, 2012.

[30] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Proc. FOCS*, 2006.

[31] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011.

[32] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized product quantization for approximate nearest neighbor search," in *Proc. CVPR*, 2013.

[33] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *NIPS 22*, 2009.

[34] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1092–1104, 2012.

[35] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *NIPS 21*, 2008.

[36] Y. Weiss, R. Fergus, and A. Torralba, "Multidimensional spectral hashing," in *Proc. ECCV*, 2012.

[37] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval,"

IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2916–2929, 2013.

[38] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. ICML*, 2011.

[39] Y. Mu, J. Shen, and S. Yan, "Weakly-supervised hashing in kernel space," in *Proc. CVPR*, 2010.

[40] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[41] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *NIPS 22*, 2009.

[42] M. Norouzi and D. M. Blei, "Minimal loss hashing for compact binary codes," in *Proc. ICML*, 2011.

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

[44] W. Kong and W.-J. Li, "Isotropic hashing," in *NIPS 25*, 2012.

[45] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. ICCV*, 2003.

[46] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[47] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[48] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[49] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[50] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[51] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[52] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[53] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[54] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in neural information processing systems*, 2004, pp. 1601–1608.

[55] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 977–986.

[56] X. Chang, F. Nie, Z. Ma, and Y. Yang, "A convex formulation for spectral shrunk clustering," in *AAAI*, 2015.

[57] Y. Li, J. Huang, and W. Liu, "Scalable sequential spectral clustering," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[58] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 214–225, 2004.

[59] H. Shinnou and M. Sasaki, "Spectral clustering for a large data set by reducing the similarity matrix size." in *LREC*, 2008.

[60] T. Sakai and A. Imiya, "Fast spectral clustering with random projection and sampling," in *Machine Learning and Data Mining in Pattern Recognition*. Springer, 2009, pp. 372–384.

[61] D. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 907–916.

[62] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 3, pp. 568–586, 2011.

[63] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation," in *AAAI*, 2011.

[64] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[65] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[66] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1977–1984.

[67] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *Neural Networks, IEEE Transactions on*, vol. 22, no. 11, pp. 1796–1808, 2011.

[68] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," *Advances in neural information processing systems*, vol. 16, pp. 177–184, 2004.

[69] A. Passerini, M. Pontil, and P. Frasconi, "New results on error correcting output codes of kernel machines," *Neural Networks, IEEE Transactions on*, vol. 15, no. 1, pp. 45–54, 2004.

[70] C. Alzate and J. A. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel pca," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 335–347, 2010.

[71] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *arXiv preprint cs/9501101*, 1995.

[72] W. Liu, J. He, and S.-F. Chang, "Large graph construction for scalable semi-supervised learning," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 679–686.

[73] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. of SDM*, vol. 13. SIAM, 2013, pp. 252–260.

[74] A. Frank, A. Asuncion *et al.*, "Uci machine learning repository," 2010.

[75] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[76] D. Dueck and B. J. Frey, "Non-metric affinity propagation for unsupervised image categorization," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* IEEE, 2007, pp. 1–8.

[77] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[78] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval.* Cambridge university press Cambridge, 2008, vol. 1.

[79] Y. Li, C. Chen, X. Huang, and J. Huang, "Instrument tracking via online learning in retinal microsurgery," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014.* Springer, 2014, pp. 464–471.

[80] Z. Pezzementi, S. Voros, and G. D. Hager, "Articulated object tracking by rendering consistent appearance parts," in *IEEE International Conference on Robotics and Automation. ICRA'09.*, 2009, pp. 3940–3947.

[81] R. Sznitman, A. Basu, R. Richa, J. Handa, P. Gehlbach, R. H. Taylor, B. Jedynak, and G. D. Hager, "Unified detection and tracking in retinal microsurgery," in *Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 1-8.* Springer, Heidelberg, 2011.

[82] D. Burschka, J. J. Corso, M. Dewan, W. Lau, M. Li, H. Lin, P. Marayong, N. Ramey, G. D. Hager, B. Hoffman *et al.*, "Navigating inner space: 3-d assistance for minimally invasive surgery," *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 5–26, 2005.

[83] R. Richa, M. Balicki, E. Meisner, R. Sznitman, R. Taylor, and G. Hager, "Visual tracking of surgical tools for proximity detection in retinal surgery," in *Taylor,*

R., Yang, G.Z., (eds.): IPCAI 2011, LNCS, vol. 6689. Springer, Heidelberg, 2011, pp. 55–66.

[84] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager, and P. Fua, "Data-driven visual tracking in retinal microsurgery," in *MICCAI.* Springer, 2012, pp. 568–575.

[85] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting." in *BMVC*, vol. 1, no. 5, 2006, p. 6.

[86] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition (CVPR).* IEEE, 2009, pp. 983–990.

[87] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection," in *Computer Vision and Pattern Recognition (CVPR).* IEEE, 2011, pp. 1313–1320.

[88] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski, "Robust and fast collaborative tracking with two stage sparse optimization," in *Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 624–637.* Springer, Heidelberg, 2010.

[89] R. Sznitman, R. Richa, R. H. Taylor, B. Jedynak, and G. D. Hager, "Unified detection and tracking of instruments during retinal microsurgery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1263–1273, 2013.

[90] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[91] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 2756–2759.

[92] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[93] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2001, pp. I–511.

[94] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.

[95] M. R. Pickering, A. A. Muhit, J. M. Scarvell, and P. N. Smith, "A new multi-modal similarity measure for fast gradient-based 2d-3d image registration," in *IEEE Engineering in Medicine and Biology Society. EMBC'09.*, 2009, pp. 5821–5824.

[96] S. Benhimane and E. Malis, "Homography-based 2d visual tracking and servoing," *The International Journal of Robotics Research*, vol. 26, no. 7, pp. 661–676, 2007.

BIOGRAPHICAL STATEMENT

Yeqing Li received his Ph.D. in Computer Science and Engineering from the University of Texas at Arlington at 2015. Prior to beginning the Ph.D. program, Yeqing obtained his B.S. degree from Shantou University, China in 2006 and his M.S. degree from Nanjing University, China in 2009, all in Computer Science. His main research interests are computer vision, image search, machine learning, and medical imaging, especially solving machine learning and computer vision problems in large-scale data set. During his Ph.D. program, he has published several papers in the top tier conferences in the literature such as the Annual International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), AAAI Conference on Artificial Intelligence (AAAI), IEEE Conference on Computer Vision and Pattern Recognition (CVPR).