DATA SYNTHESIS FOR OBJECT RECOGNITION

BY

XI ZHANG

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science
in the Graduate College of the
Illinois Institute of Technology

Approved _____
Advisor

Chicago, Illinois
December 2017

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Large and balanced datasets are normally crucial for many machine learning models, especially when the problem is defined in a high dimensional space due to high complexity. In real-world applications, it is usually very hard and/or expensive to obtain adequate amounts of labeled data, even with the help of crowd-sourcing. To address these problems, a possible approach is to create synthetic data and use it for training. This approach has been applied in many application areas of computer vision including document recognition, object retrieval, and object classification. While a boosted performance has been demonstrated using synthetic data, the boosted performance is limited by two main factors in existing approaches. First, most existing approaches for creating and using synthetic data are application-specific and thus lack the ability to benefit other application areas. Further, such application specific approaches are often heuristic in nature. Second, existing approaches do not recognize an inherent difference between synthetic data and actual data which is termed as a synthetic gap in my proposal. The synthetic gap in existing approaches is due to the fact that not all possible patterns and structures of actual data are present in the synthetic data. To address the problems of using synthetic data and using it to better improve the performance of learning algorithm, this proposal considers general ways of creating and using synthetic data. The problem caused by the synthetic gap is studied and approaches to overcome the gap are proposed. Initial results demonstrate that the proposed approach is effective and can boost the performance to many computer vision applications including building roof classification, character classification, and point cloud object classification.

CHAPTER 1

INTRODUCTION

1.1 Overview

Computer vision is a field that focus on methods for acquiring, processing, analysing high-dimensional data from the real world, thus recognize and understand the incoming information. The recognition is based on transforming these input data to symbolic information, e.g., via decisions[**?**, **?**, **?**, **?**]. A primary mission of computer vision researchers is to simulate or improve the abilities of the human vision system by processing and understanding incoming visual data. The understanding of data in computer vision research can be viewed as processing of detangling of symbolic and context information from incoming visual signals using methods and models constructed with the aid of computational methods from other disciplines, such as: geometry, physics, statistics, and artificial intelligence[**?**].

Recognition based on computer vision is an interesting and challenging topic that has attracted many research efforts all around the world. From simplest geometry pattern recognition to complex scene understanding based on statistical theory and artificial intelligence, the techniques used in computer vision recognition are becoming more and more advanced. In the same time, the number of applications involving computer based recognition system has expanded quickly in recent years. Nowadays, searching applications of computer vision on the internet, will return hundreds of categories that are built upon computer vision, such as: face recognition, autonomous cars, gesture recognition and so on.

Development in recent years such as deep neural networks(DNN)[**?**], deep belief networks(DBN)[**?**], convolutional nerual networks(CNN)[**?**] and convolutional deep belief networks(CDBN)[**?**], contribute to many successes in visual recognition tasks.

To mine and understand the intrinsic relation and correlation of the data, researchers

have made many efforts for designing high dimensional data features that can better characterize the nature of the data. Generally, due to the complexity of the models and the dimensions of the features, a huge number of training data is required[**?**, **?**]. A learning process which is short of enough training data could result in many problems. One of most well-known one is over fitting where the resulting model perfectly fitted to a few training data and thus the model has large bias as with respect to the underlying ideal model. It has been proven that with a fixed number of design pattern samples, the accuracies of the model reduces as the dimensionality increases[**?**], the problem of which is known as the curse of dimensionality.

Data is extremely valuable and expensive. On one hand, the process of the data acquisition is very expensive and may require human labor and resources. For example, to collect real-world street view data for the computer vision benchmarks KITTI, an experienced car drive has to drive a standard station wagon through every streets in the city which equipped with two high-resolution color, a gray scale video cameras, a GPS localization system and a Velodyne laser scanner[**?**]. The expense of the data acquisition could go even higher when more costly data has to be collected. For example, in remote sensing, to have a detailed 3D mapping of a terrain of a region, an aircraft has to be sent with an experienced pilot and a complete set of expensive and professional remote sensing equipment. On the other hand, data labelling is much more expensive than data acquisition, and sometimes it is even impossible to get data labelled. For example, when a medical diagnosis has to be made by a model learned using hundreds of thousands of medical images, to guarantee the accuracy of prediction is in a confidence interval that is under control, the labelling of data has to be operated under advices of the physicians. Another example in which labelling is almost impossible is the recognition of the point cloud objects from street view point cloud data that is collected from real-world. A snapshot of such data is shown in Figure **??**. We employ a learning based framework in this application, objects are categorized to 6 classes: car, pedestrian, street-light, traffic-light, tree and trash can. To have a well labelled training

set, we spend numerous hours on manipulating the data in 3D and selecting and assigning labels.

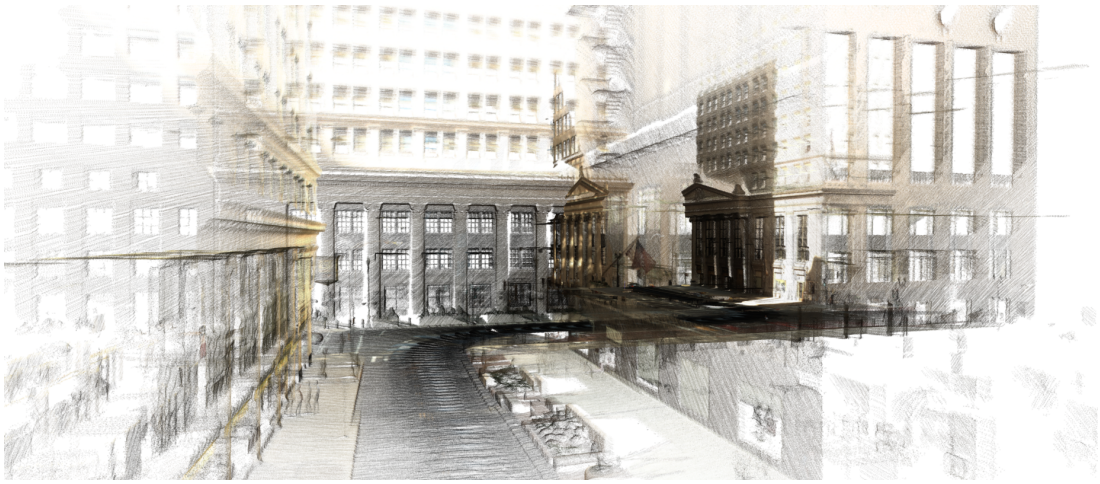


Figure 1.1. A screen-shot of street view point cloud data used in my research. There are 9195492 points that are contained in this scene.

1.2 The Proposed Solutions

Data labelling is a tedious and time consuming step of a machine learning process, and may be very expensive especially if experts are involved. To tackle this problem and reduce the cost spent on data labelling, several strategies and techniques had been proposed in last two decades.

A technique that creates synthesized data by over-sampling existing data has been extensively used in computer vision and machine learning. The most popular one among these techniques is SMOTE [**?**] which over-samples a minority class by using interpolation in feature space. Such interpolation gives a larger and more confident decision region to minority classes. Hence, the learner learns a less-biased boundary between the classes.

Instead of creating new synthetic data, a technique called semi-supervised learning will start from a small amount of labelled data and gradually grow and propagate to get more data labelled. In order to use unlabelled data, several assumption has to be made to the underlying distribution of data, such as: a smoothness assumption, a cluster assumption

and a manifold assumption[**?**]. Under this configuration, semi-supervised learning attempts to use this combination on labelled and unlabelled data were made. The idea is to improve the classification performance by propagating the label of labelled data to unlabelled data so as to expand the set of the data with labels.

All of the learning schemas and techniques mentioned above have been successfully applied by researchers in the of computer vision and machine learning for decades. However, almost all of these methods are based on an assumption that the distribution of the data is known. Such assumption usually fails unless either a data set with distribution that covers every possibility in feature space is given, which push us back to the problem of data acquisition, or the data of training set and testing set are exactly drawn from the same distribution. An example that can explain this phenomenon is given by the problem of recognizing the human body motion. To have a high accuracy recognition of human body motion, a human body model with high order of degrees of freedom is assumed with each degree of freedom representing a dimension in movement space. If the body movements in training set and testing set are totally different, (say training set containing arm movement only but testing set containing head tuning only), then the learned model has little or even no chance to recognize the testing set.

In many recognition problems, humans have prior knowledges regarding the structure and the parameter domain of the problem. Take the problem of roof style recognition from satellite images for example, the goal in this problem is to recognize the roof style of a given roof satellite image. The target labels are: hip, halfhip, gable, pyramid, gambrel, and flat. The recognition of the roof style usually is conducted by recognizing features such as ridge lines and valley lines on the roof. However, due to the degradation of image quality and occlusion caused by shadows, it is hard to detect all of these features, the cases of which are shown in Figure **??**. To solve this problem, since we know how roof is built and we know where those ridge lines and valley lines are located on the roof, hundreds of

thousands of synthetic roofs containing these features could be created and then used to train a recognizer.

Examples of using synthetic data to assist recognition process are very common. The success of these methods and applications lie in the fact that on the one hand, the essential features of an object could be captured by a synthetic model using low level object features, such as the object silhouette, object color and object structure. Thus training a recognizer using synthetic data could capture low level features of objects and can assist the recognition process. A well known application of using a synthetic model for recognition is an object retrieval system in which users are required to provide a rough drawn sketch of underlying target[**?**], where the object sketches drawn by the user provide both the silhouette and shape information of an object.

A key problem of using synthetic data for recognition is the mapping between real data and synthetic ones. Reasons for such data mapping come from intrinsic differences between real data and synthetic data. Consider the roof image classification for example, the synthetic roof images are modelled as a parametric model of the combination of the roof's ridges, valleys and corners. By tuning parameters of these elements, we can generate a huge number of synthetic roofs, However by creating synthetic data, it is impossible to simulate numerous possibilities of noisy patterns in real roof images. Therefore mapping is needed to eliminate the gap between real data and synthetic ones. The mapping can be done in many ways and in different stages of the data processing. For the roof case, given a real roof, we need to match it with a synthetic roof so that the different between two is minimal. In addition we need to employ some transformation in feature space to "add" noisy patterns to synthetic data or 'remove" noisy patterns from real data. In this research, I deal with the mapping from both data space and feature space. In data space, I proposed to build a synthetic model that will minimize the distance between real data and synthetic ones. In feature space, since the distribution of the synthetic data is different from the distribution

of real data, I consider method that can make a shift for the features of either real data or synthetic data to reduce disparities between the two data. To successfully achieve these objectives I use different computer vision and artificial intelligence techniques, such as: deep learning and transfer learning.

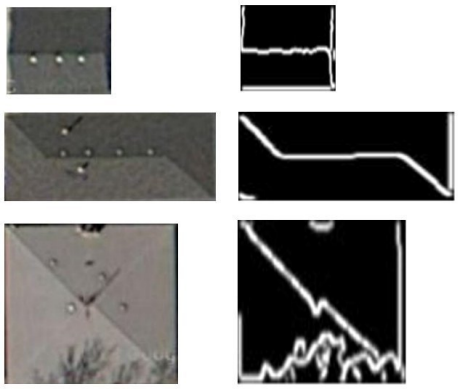


Figure 1.2. A demonstration of using synthetic template to simulate edge image of roofs used in my research.

1.3 Novel Contribution

In this researches, I address recognition problems of 2D & 3D visual data using learning based methods. Generally there are two factors determining the result of a learning process.

The first factor is the learning algorithm. A successful learning algorithm should be effective and robust to unseen data. The second factor determining the learning result is the data. Data preparation usually is a pre-processing step in machine learning procedures and is often not given enough attention relatively. The quantity and the quality of the data are key factors affecting the success of a learning procedure.

It is generally accepted that the performance of a learning process could be increased if more data is available. However, data is always not enough. In my research, to solve the lack of data problem, I proposed to use synthetic data to help increase the performance of a learning process. There are various problems when it comes to data synthesis in

a machine learning procedure. To sum up, there are two major problems we need to solve. The first problem we has to solve is where to synthesize more data. The word ”where” in this context has two meanings. We first should determine in what space to conduct such data synthesis, in data space or feature space. Then we need to determine the location in each space where to add the synthetic data. These are very important questions we should really take care of, because answers to any questions will determine the complexity of the data synthesis process in the first place and the performance of the entire learning process at the end. The second problems deals with how to use the synthetic data in a learning procedure. To simulate actual data in a learning process, the synthesized data has to be as similar to actual data as possible. The synthetic data is useless unless it can truly mimic actual data. To solve these problems, my research has the following primary contributions.

1.3.1 Creation of Synthetic Data in Data Space. Machine learning algorithms usually suffers from lacking of enough training data. In existing methods researchers usually use a technique called data augmentation to increase the amount of training data. In data augmentation, new examples are generated as a perturbed version of the original data. Take image augmentation for example, to augment new images, the original images are stretched, added noises or partially cropped from the original images. Images generated in this way form a distribution in parameter and feature space which is centered at the original images. Thus, these new images do not introduce much new informations in this case. Therefore, the learner may fails to learn a more general decision boundary from the resulting data and the increase in performance will be limited.

In my research, to better simulate actual data, a series of methods that synthesize more data are introduced. These methods can be basically separated to two groups. In first group, different from existing methods, my methods of data synthesis are based on usage of a centralized representation of actual data. We call such centralized representation a prototype in this proposal, because the prototype can characterize the most essential structures

and patterns of actual data. The prototype in my work is modelled as a parametric model of control points and could be learned using semi-automatic or automatic approaches. Using technique such as interpolation between existing data and randomly drawing synthetic data from a distribution computed from control points of existing data, more data could be synthesized easily. Methods in second group are machine learning based methods, along the recent year' development of deep neural network, especially in computer vision area. Large dataset has become a necessary input to neural networks to gaurantee a desired performence of the neural networks. Image level or even pixel level data synthesis become even valuable and important in these tasks. In my work, depending on learning and computational power of deep neural networks, I explored and proposed ways that synthsize image level data by letting neural network learn existing data. The synthesized data whose usage is demenstrated in application such as data retrievel, object tracking and so on.

The proposed approach has been used in many object recognition problems in my research. The preliminary results can be found in [**?**][**?**][**?**].

1.3.2 Learning From Synthetic Data. Directly using and learning from synthetic data may cause failure or a classification results with poor performance, even though there is enough synthetic data. This is because, the synthetic data is generated using a data prototype(parametric model) which is a simplified version of data representation. Since there is no way for a prototype to generate information such as noise and some other complicated patterns of actual data, this results in nonequivalence in terms of information contained by features computed from actual data and synthetic data. Thus synthetic data holds a different decision region with respect to actual data and the result may be poor classification performance if learned directly from these data.

A general goal of feature extraction is to minimize the difference between actual data and synthetic data by using extracted features. To achieve this goal, I propose two strategies in feature extraction that prove to be useful. First, extracted features have to be

able to capture the most essential characteristics between actual data and synthetic data, thus ignoring noise all the noises and inherent patterns. Second, extracted features should be able to compensate for such additional information for synthetic data. Under these two directions, several techniques and algorithms are developed in this work.

The preliminary results of the proposed feature extraction strategies could be found in [**?**][**?**][**?**].

1.3.3 Elimination of Synthetic Gap. Learning a classifier from synthetic data is extremely hard, due to the following reasons. First, using a prototype to synthesize more data causes the distribution of the generated data to shift away from that of actual data. We term such distribution shift as *synthetic gap* in this proposal. The *synthetic gap* is a major obstacle in learning from synthetic data, since synthetic data may fail to simulate the potential useful patterns of real data for training classifiers. Second, since practically a small amount of labelled data may be available, it is necessary to jointly learn from synthetic and real data. The learning process must be automatically leveraged between the synthetic data and the real data.

Different from existing work, here we solve the problem from a more general perspective by eliminating the *synthetic gap* as a domain adaptation problem. A general framework based on a neural network with novel architecture is proposed. The neural network is trained as a regressor which could map synthetic data to real data and find out the potential transformation between two types of data. Thus it is able to close the gap between two data. With this mechanism synthetic data could better simulate actual data and used in a machine learning procedure.

1.3.4 Creation of Synthetic Data in Feature Space. Quality of datasets play an important role in machine learning. Somebody even say the dataset is the root of a successful machine learning method. However, in practice, many applications and machine leanring problems

suffer from degraded performance due to unbalanced distribution of datasets.

Data synthesis through prototype can be used to improve performance of many applications suffering from unbalanced data distribution. However, due to the nature of the methodology, it is limited to some specific problems. A more general solution is to directly create synthetic data representation to achieve a balance in its feature space.

Synthesizing new data in feature space is hard. Because data representation in feature space is abstract that the actual meaning of the data in data space is hard to interpret. To know where to synthesize new data so as to balance the data distribution and improve the performance of underlying machine learning tasks such as classification, it requires a deep understanding of distribution of entire dataset. To achieve this goal, I proposed a theorotically gauranteed minority oversampling algorithm which takes both data distribution of minority and majority class into account and can synthesize data to achive better classification results compared to existing methods. Details of the algorithm could be found in Chapter **??**.

CHAPTER 2

CREATION OF SYNTHETIC DATA IN DATA SPACE

2.1 Introduction

It is generally accepted that an accurate classifier can be learned if a large and balanced labeled training dataset is available. In real-world scenarios, however, one always struggles to find adequate amounts of labeled data. Even with the help of crowdsourcing, e.g., Amazon Mechanical Turk (AMT), it is often difficult to collect a large quantity of labeled instances with high quality that is necessary for training a classifier for a real-world problem. In terms of quantity, it has been shown that amount of available training data, per class, follows Zipf distribution [**?**]. In terms of quality, some domains, such as analysis of satellite images(e.g. the comet images from Rosetta), require extensive and detailed expert user annotation [**?**]. Large volume of LiDAR point cloud data have to be labeled before they can be used to train some classifiers [**?**]. Such labeling process usually is very time consuming and requires expert-level labeling efforts or expensive equipments. Practically only a very limited portion of the data points can be obtained. The fundamental issue of learning from not enough data is the ability of these data to significantly compromise the performance of most standard machine learning algorithms. Most standard learning algorithm, generative and discriminative, assume that enough knowledges and features could be extracted from training data that have a balanced distribution. Therefore, when presented with a complex dataset but with less amount of data, these algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable accuracies.

Technically, there is no official definition or a quantitative measure about rare data learning. It is important to understand rarity of different forms. According to [**?**], rarity could comes from either relative rarity which is usually related to imbalanced learning or absolute rarity which is a direct result of the nature of the data set. my method and solution partially overlap with imbalanced learning in cases when data is imbalanced distributed,

and the minority classes are viewed as rare data. More than rare data learning, methods discussed in this chapter solve more general problems in which labelling actual data maybe is very time consuming or a job must need a lot of labor works to be done.

I propose to create synthetic data and use them to solve mentioned problems in a machine learning process. The idea of using synthetic data in machine learning algorithm has a long history and is associated with the development of cognitive psychology, artificial intelligence and computer vision. The reason simply lies in the facts that first creating synthetic data is much easier when actual data could be represented as a parametric model. Second, if I view a class of objects in data space as a whole, synthetic data and actual data are both subsets of the entire dataset which may or may not overlap. Indeed synthetic data is no more than an ideal representation of actual data. As an example, I show edge images extracted from my satellite roof top images and its corresponding synthetic images in Figure **??**. It could be observe from this figure that actual images are highly similar to synthetic ones and could be identical to synthetic ones if noises can be removed.

2.2 Related Work

Due to technical constraint in five years, data synthesis is limited to just using geometrical transformation and degradation models. In [**?**][**?**], to help off-line recognition of handwritten text, a perturbation model combined with morphological operation is applied to real data. They showed that when a moderate transformation is added to the real data, the resulting synthetic training set boost the performance. Synthetic data generation techniques has been moderately implemented in area of optical character recognition. in [**?**], researches and experiments were conducted to test the safeness of using synthetic data in machine print text recognition problem. Following the conventional way of generating synthetic data, they model each synthetic data as a geometrically transformed version with different magnitude of the original data. They demonstrated that classifier trained on interpolated data often but not always improved classification when tested on previously unseen

samples. However, the use of interpolated data in the training sets has never worsened the results. To enhance the quality of degraded document, in [**?**] degradation models such as brightness degradation, blurring degradation, noise degradation and texture-blending degradation were used to create a training dataset for a handwritten text recognition problem. The synthetic minority oversampling technique (SMOTE) [**?**] and its variants [**?**][**?**] are also powerful methods that have shown many success in various applications. However, these previous methods are relatively limited to one particular type of dataset, whilst I propose a more general methodology of generating synthetic data by creating a parametric prototype of the data. By tuning parameters of the prototype, my method can derive synthetic data that cover almost all possibilities. A very interesting work whose intention is a bit similar to mine is proposed in [**?**]. In this paper, to analyse variability in point-sampled geometry, authors first assume every data comes in with incomplete information about ground true object. They then capture this uncertainty by introducing a statistical representation that quantifies for each point in space the likelihood that a surface fitting the data passes through that point. This likelihood map is constructed by aggregating local linear extrapolators computed from weighted least squares fits. The quality of fit of these extrapolators is combined into a corresponding confidence map that measures the quality of local tangent estimates.

In recent five years, along with the development of deep learning techniques. Research of data synthesis, particularly image synthesis, has receive more and more attention. There are two reasons for this phenomenon. On the one hand, large deep learning system usually need large amount data to gaurantee a unbiased and high quality model (network/networks), which make desire for synthetic data become more imperative. On the other hand, modern deep learning techniques provide a efficient and prominent way to synthesize data.

Many efforts have explored using synthetic data for various prediction tasks, includ-

ing gaze estimation [**?**], text detection and classification in RGB images [**?**], font recognition [**?**], object detection [**?**], hand pose estimation in depth images [**?**][**?**], scene recognition in RGB-D , semantic segmentation of urban scenes [**?**], and human pose estimation [**?**][**?**][**?**][**?**][**?**][**?**]. Gaidon [**?**] show that pre-training a deep neural network on synthetic data leads to improved performance.

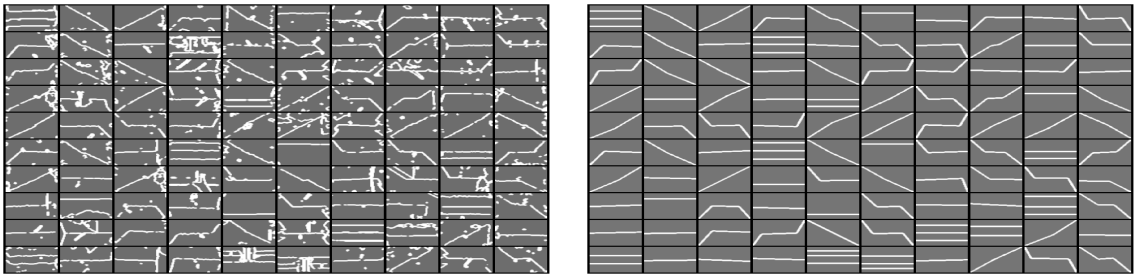
Due to a technique called Generative Adversarial Network (GAN) introduced by [**?**], Generating complicated deceptive artificial images become truly possible. Basically, the GAN framework learns two networks (a generator and a discriminator) using competing loses. The goal of generator network is to map a random noise to realistic images so as to fool discriminator network, whereas the goal of the discriminator network is to distinguish the generated from the real images. Since introduction of GAN, many extended works have been proposed for data synthesis purpose [**?**]. Based on several improvements on basical GAN network structure, Radford [**?**] proposed a network framework that could synthesize much more realistic images. Li and Wand [**?**] proposed a Markovian GAN for efficient texture synthesis. Lotter use adversarial loss in a LSTM network for visual sequence prediction. With auxiliary classifiers added to GAN model, [**?**] investigate performance of variants of GAN models on image synthesis task. By using two GAN models in a cycle framework, cycleGAN [**?**] is able to acheive image style transfer. Later, a conditional cycleGAN is introduced in [**?**] for more specific image synthesis tasks. In my work, I proposed to synthesize realistic image using image descriptions. The earlist work of this task could be back to [**?**], although no GAN is used in this work, by using recurrent neural network, this work found the basic framwork of text to image transfer. Combining GAN and text features generated from a LSTM, [**?**] could parse image description to synthesize realistic image corresponding to the description. The same idea is found in [**?**], however, LSTM is co-trained with GAN in this work. A stackedGAN structure is proposed in [**?**], due to image quality refinement broght by second GAN in this stacked-up structure, the network is able to synthesize highly realistic images.

2.3 A Brief Introduction of Proposed Methods

In my approaches, I create synthetic data in data space. Creating synthetic data in data space has several advantages. First, it is the most intuitive way to create new data, because it is easy to judge the quality of the new data. Second, creating data in data space let us understand my data more clearly and deeply, because in order to create new data, I have to understand the characteristics and structure of original data. Having these understanding, it is more likely I will invent some better features that would boost the performance of machine learning process later on. The feasibility of my method is based on an observation that data of most problems can be converted to a parametric model, in my cases 2D and 3D models controlled by multiple control points. By adjust values of parameters I are able to derive many synthetic data.

In my research, the approach of creating synthetic data in data space primarily solves two types of problems. First, by creating more synthetic data, I solved problems of rarity learning. Because as I mentioned that both actual data and synthetic data are subset of entire dataset. By creating massive number of synthetic data, the approach enriches and compensates the part in data space where actual data does not reach, so that a more complete distribution of the data and a more tight boundary between classes of data are provided to machine learning algorithm to learn. Second, for some problems, using synthetic data will greatly reduce labor works spending on labelling actual data. Because instead of label huge number of actual data, using my method, by indicating a label of data, I can generate as many as possible synthetic data I need. For these two problems, I will use the rest of this chapter to discuss the details of creating synthetic data in my previous works. Also, I am going to talk about the problems and challenges in my future research.at the end of this chapter.

2.4 Creating Synthetic Data to Avoid Rarity Learning



(a) Real roof edges (b) Synthetic roof edges

Figure 2.1. Examples of (a) real roof edge vs. corresponding (b) synthetic roof edge images. The synthetic data is generated by the algorithms in Sec. 4. The examples are randomly drawn from SRC dataset.

Learning from rare data is very challenging, because rare data are typically much harder to identity than common data and most standard machine learning algorithms have a great deal of difficulty to detect regularities within the rare data.

Generally, there are two types of rarities in context, Much of the research on rarity relates to *rare classes*, or more generally, class imbalance. This type of rarity requires labeled examples and is associated with classification problems. A second type of rarity concerns *rare cases*. Informally, rare cases correspond to a meaningful but relatively small subset of the data, or equivalently, define a small region of the instance space. Rare cases depend only on the distribution of data and therefore are defined for both labeled and unlabeled data, for supervised and unsupervised data mining tasks. Rare cases are naturally defined by the domain and will share common characteristics. In the case of labeled data, a rare case corresponds to subconcept, or subclass, that occurs infrequently.

There are a number of problems that arise when mining rare classes and rare cases. The first problem I could encounter is lack of data. In this case, the number of examples associated with the rare class is small in an absolute sense which make it difficult to detect regularities within the rare classes/cases. Another problem could that could happen is generating a learning model by using an inappropriate inductive bias. Generalizing from specific examples, or induction, requires an extra-evidentiary bias. Without such a bias

”inductive leaps” are not possible and learning cannot occur. The bias of a data mining systems is therefore critical to its performance. One more difficulty comes with noise in training data, especially when rare classes/cases are associated with noise. Since noise data will affect the way any data mining system behaves, but what is more interesting in my research is that noise has greater impact on rare cases that on common cases. Because rare classes have fewer examples to begin with, it will take fewer ”noisy” examples to impact learned subconcept. In this case, the learner cannot distinguish between exceptional (rare) cases and noise.

Among all of these challenges and difficulties, in my research, I primarily deal with supervised classification problem which may have *rare classes* or *rare cases* or both. By synthesize more data to create a much more balanced training set using my techniques, I demonstrate that a better performance could be obtained in many classification problems.

2.4.1 Satellite Image Roof Style Classification. Recognizing building roof styles is an important step in many applications including building reconstruction and urban planning. Such problems generally need very high quality training data. However, there is no previous dataset for such problem. Usually collecting data through online crowdsourcing, e.g., Amazon Mechanical Turk (AMT) is very expensive. So to facilitate the study, in my works [**?**][**?**], a few volunteers are invited to join the project and manually crop images from civilian level online digital maps. Manually cropping images is an tedious and time consuming task. By the end of data collection, I are able to only crop a few thousands of roof images.

For a large building which is usually comprise of multiple parts, I even develop an algorithm that reasonably segment large roof image to simple parts. I design an algorithm to decompose complex footprint. The basic idea of this algorithm is to segment the concave footprint into several convex polygons, while the cutting line should make the new polygons have less sum of variances of all inner angles in them. The reason I want to use convex polygon rather than concave polygon to recognize is generally the roof styles can

Algorithm 1 Footprint Segmentation

Input:
The set of footprint vertices, V;
Output:
The set of selected segmented region vertices sets, S;

```
Initial S;
PUSH(V, S);
if |V| > 3 then
    for i = 1 to |V| do
        if V_i is POI then
            Calculate cutting line {V_i, D}, where D is the other endpoint;
            PUSH ({V_i, D}, C);
        end if
    end for
    for i = 1 to |C| do
        for j = 1 to |S| do
            POP (R, S), where R is the top element in S;
            Cut R by cutting line C_i into R_1 and R_2;
            PUSH (R_1, S);
            PUSH (R_2, S);
        end for
    end for
end if
return S;
```

be built on convex polygon much easier.

To achieve this, I assume the vertex whose inner angle is larger than as point of interested (POI). Then I need to traverse all possible angles from POI and find which cutting line leads the minimum variance of all inner angles in this segmentation. With calculated cutting lines and given footprint vertices, I can segment the polygon into basic parts. The algorithm shows in Algorithm **??**, Let V be the set of footprint vertices, which elements has been sorted in clockwise or counterclockwise, S be the set of selected segmented region vertices sets while S is a queue. C is the set of cutting lines. R is a set of polygon vertices and R_1, R_2 are the segmented polygon vertices.

The cutting line calculation follows:

$$\sigma(i) = \sum_{j=1}^{2} \sum_{k=1}^{|A_j P|} [A_{jP(i)}(k) - \frac{(|A_j P| - 1) \times \pi}{|A_j P|}] \tag{2.1}$$

Where i is a given cutting line, $P(i)$ is the set polygons which generated by cutting line i. $A_j P$ is the set of inner angles of $j^t h$ polygon in set P. Hence the appropriate cutting line should be $\arg\min(\sigma)$, while $\sigma(i)$ is the set of variance of cutting line i. The time complexities of cutting line calculation and polygon segmentation are constants. Hence the total theoretical maximum time complexity is $(|V| + 2^{|C|})$, while $2^{|C|}$ always has same magnitude with $|V|$.

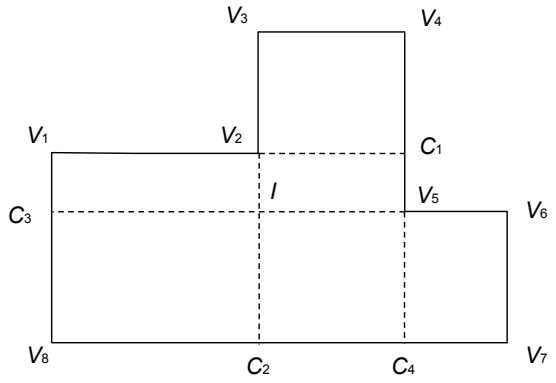
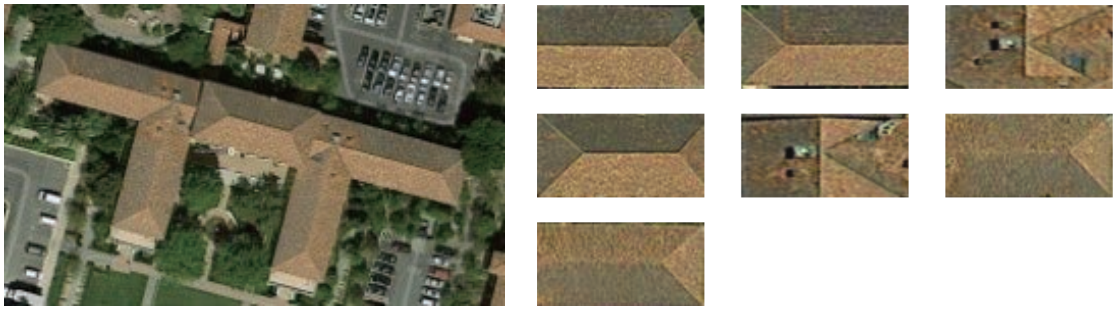


Figure 2.2. An example of input polygon vertices, cutting lines and six segmented components result.

A detailed example is shown in Figure **??**, with input footprint vertices V. The first loop in algorithm can figure out $\{V_2, C_1\}$, $\{V_2, C_2\}$, $\{V_5, C_3\}$ and $\{V_5, C_4\}$ are cutting lines. The first cut to V by $\{V_2, C_1\}$ will get $\{V_2, V_3, V_4, V_5\}$ and $\{V_1, C_1, V_5, V_6, V_7, V_8\}$ these two polygons. Iterate all cutting lines until I get six segmented polygons.

An example of roof segmentation using the proposed approach is given in Figure **??** in which a building roof is divided to 7 pieces.

Later all roof top images are aligned using an approach proposed in [**?**]. Then all



(a) (b)

Figure 2.3. An example of segmentation on a complex building roof. All pieces are normalize to a same dimension.

images are resized to the same dimension. This dataset is of great challenges for the task of visual analysis. First, qualities of the some satellite images are degraded because of significant image blurring occurred when capturing the satellite images. Second, roofs in these images are covered by various kinds of equipments such as air conditioners chimneys and water tanks, and most of roofs in my dataset are partially occluded by shadows cast by trees and some other stuffs. Such covering and shadows are great obstacles to robust visual analysis algorithms. Examples of my data in this dataset is given in Figure **??**

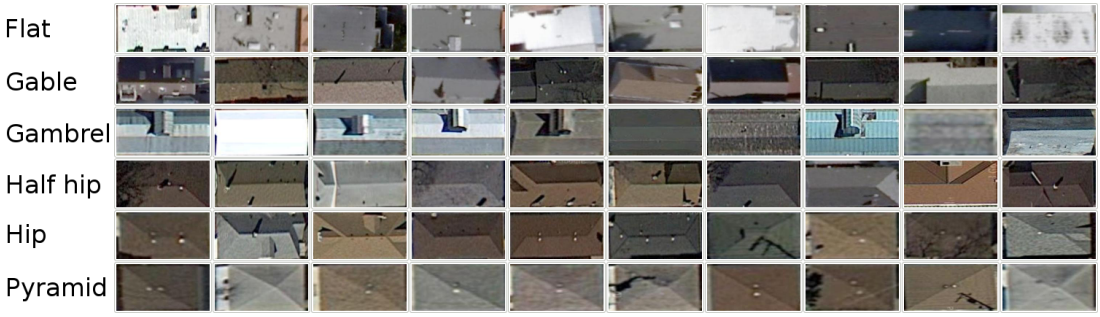


Figure 2.4. Examples of roof top images used in my work.

Furthermore, the class instances of this dataset are intrinsically extremely unbalanced that some particular types of roofs (such as gambrel and pyramid) are far less than the other types. Such unbalanced distributions of data are compared in Table **??**.

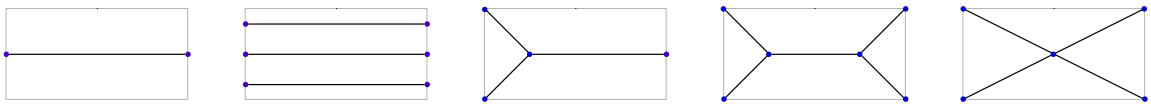
Classification of the roof styles in the experiments are based on recognizing edges detected from the roof images. I employed the adaptive Otsu edge detection method [**?**]

Table 2.1. The distribution of the roof styles used in the experiments.

Styles	Training #	Testing #	Total #
Flat	1232	1748	3080
Gable	1111	1665	2776
Gambrel	156	232	388
Halfhip	268	400	668
Hip	960	1440	2400
Pyramid	133	199	332

to extract edges from the roof images. Examples of generated edge images are shown in **??** (a). The synthetic prototypes are then created to characterize primary structures of a roof represented by extracted edges. The way I create synthetic models is to simplify and convert roof edge images to a parametric model in which edges are connected by a few control points. By adjusting the position of the control points and drawing lines between the points, I are able to simulate different variations of roof edge images. All noises and line patterns introduced by actual roof edge images are ignored in synthetic images, which makes it easier to create synthetic images. However, without present these noises and patterns in synthetic images potentially increase the distance between actual data and synthetic data, thus lower the accuracy of classifier trained using synthetic data. I will talk about techniques and algorithms dealing with these issues in Chapter **??** and Chapter **??**

For all kinds of roof styles except flat I classified in my work, I intuitively design the parametric models of these roof edge images as the ones shown in Figure **??**.



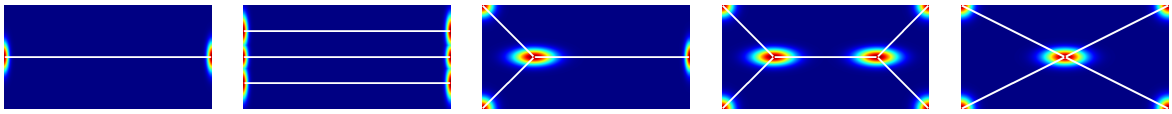
Gable Gambrel Halfhip Hip Pyramid

Figure 2.5. For five roof styles used in my work, control points of each style are given in above figures which are shown as blue dots.

With these basic synthetic template, generate more synthetic data is as easy as

change positions of the control points. However, configuring these control points with arbitrary ranges will easily lead to a roof image that is impossible exist in real world thus introduce noises and outliers to dataset. To solve these problems, two approaches are proposed in my work [**?**] and [**?**] separately.

In [**?**], I assume the position of each control points in image follows a normal distribution that the set of control points: $\mathbf{P} = \{p_i\}_{i=1}^{k}$, $p_i \sim N(\mu_i, \sigma_i)$. Thus for all control points $\mathbf{P}$ of a roof style, they follow a multi-variant normal distribution that $\mathbf{P} \sim N(\mu, \Sigma)$ where Σ is covariance matrix of $\mathbf{P}$. To estimate parameters mean and covariance of existing data set, volunteers are invited to help us mark the control points of actual data. Then μ and Σ are computed from the marked point positions. To illustrate distribution of each control points in my roof edge templates, I rendered points' distribution in Figure **??**. At this point, control points of a synthetic data could be randomly drawn from multivariate normal distribution $N(\mu, \Sigma)$ I just computed.



Gable Gambrel Halfhip Hip Pyramid

Figure 2.6. For five roof styles used in my work, control points of each style are given in above figures which are shown as blue dots.

There are several disadvantages in above approach by inviting volunteer to do the labelling and assuming underlining distribution of control points. First, manually marking control points is an annoying task that it is time consuming and easy to make mistakes. Second, assuming underlying distribution of the control points is multivariate normal distribution may be problematic, especially when the number of samples is small, such as gambrel and pyramid roof styles in my work.

So, in [**?**], I propose a different method to obtain positions of the control points and construct synthetic data. In [**?**], the synthetic data are represented as a parametric

model of a set of control points and edges associated to these points in the images. From the control points, the synthetic images could be generated to simulate the real images in terms of having the same structure or a similar appearance. Initially, the control points are selected from a synthetic prototype that generalize all images in the same class. Then the locations of the control points are iteratively optimized until convergence in order to minimize the distance between synthetic images generated by control points and the real image. I annotate the control points and edges associated to them as $\mathbf{S} = \{\mathbf{P}, \mathbf{E}\}$, where $\mathbf{P} = \{p_i\}_{i=1}^{n}$ is the set of the control points, and $\mathbf{E} = \{(p_i, p_j)\}, 1 \leq i, j \leq n$ is the set of edges connecting control points. A generalized algorithm of getting the best matching synthetic image is provided in Algorithm **??**.

Algorithm 2 Get Matching Synthetic Image.

Input:

- A real image U.
- A set of control points $\mathbf{S} = \{\mathbf{P}, \mathbf{E}\}$ with all control points $p_i \in \mathbf{P}$ set to their initial positions.
- A prototype image V generated using the initial **S**.

1: **while S** is not converged **do**
2: **S** = OptimizeControlPoints($U, V, \mathbf{S}$).
3: Generate V using **S**.
4: **end while**
5: Generate synthetic image I using **S**.
6: **return** I.

In Algorithm **??**, the OptimizeControlPoints(U, V, **S**) function is a process that searches for optimal control point locations which results in a synthetic image minimizing the discrepancy between the real image and the synthetic image. A coordinate descent framework is employed to accelerate the search process. I summarize this method in Algorithm **??**.

As I mentioned previously, make an assumption that the underlying distribution of control points is multivariate normal distribution may not be appropriate in some cases, especially for a case where only a few samples are given in a high dimensional space. So, I propose another technique which does not make any assumption about the distribution of

Algorithm 3 OptimizeControlPoints($U, V, \mathbf{S}$) Case 1

Input:

- A real image U.
- A prototype of the synthetic image $\mathbf{S} = \{\mathbf{P}, \mathbf{E}\}$.
- A synthetic image V generated using $\mathbf{S}$.

for $p_i \in \mathbf{P}, 1 \leq i \leq n$ **do**
  Update $\mathbf{S}$ by moving p_i by one unit.
  Generate V using $\mathbf{S}$.
  if $\mathbf{S}$ does not reduce Dist(U, V) **then**
    Cancel the last move of p_i.
    Generate V using $\mathbf{S}$.
  end if
end for
return $\mathbf{S}$.

the control points, thus no parameter need to be estimated if using this method.

Algorithm 4 RNNI(I, k, N)

Input:

- Set of input images that are to be interpolated $I = \{\mathbf{S}_i\}_{i=1}^{n}$.
- The number of nearest neighbors k.
- The number synthetic data to be generated N.

Output:

- Set of generated synthetic data I.

for $1 \leq i \leq N$ **do**
  Randomly choose a integer number idx that $1 \leq idx \leq n$.
  Randomly choose a neighbor from k nearest neighbor of $\mathbf{S}_{idx}$ and call it $\mathbf{S}_j$.
  Randomly choose a floating number r that $0 \leq r \leq 1$.
  $\mathbf{P}_{new} = \mathbf{P}_{idx} + r \times (\mathbf{P}_j - \mathbf{P}_{idx})$.
  Connect $\mathbf{P}_{new}$ in a proper way to draw roof edges $\mathbf{E}_{new}$.
  $\mathbf{S}_{new} = \{\mathbf{P}_{new}, \mathbf{E}_{new}\}$.
  $\mathbf{I} = \mathbf{I} \cup \mathbf{S}_{new}$.
end for
return $\mathbf{I}$.

The synthetic data generated in this approach is based on interpolation between a sample data and one of its nearest neighbors. The interpolation is done on each existing sample. I call this method as Random Nearest Neighbors Interpolation (RNNI) to emphasize the randomness in the selection of nearest neighbor. Similar methods are used in a lot of previous works such as [**?**], where to facilitate character recognize, interpolation is conducted on parameters of transformation such as skew, scale and so on. my method is different from previous works that instead of compute underlying transformation, for the

sake of using control points the interpolation in my approach is directly applied on data itself which makes the entire procedure easier and effective. The algorithm that generate a set of new synthetic data is given in Algorithm **??**.

2.4.2 Hand Written Digital Character Recognition. Synthetic data is helpful in a lot of applications, another example that benefit from my method is hand written digital character recognition. I also validate my framework on handwritten digits dataset from UCI machine learning repository [**?**] which totally has 5620 instances. The handwritten digits from 0 to 9 in this dataset are collected from 43 people: 30 contributed to the training set and the other 13 to the test set.

Generation of synthetic data in hand written digital character dataset is still done by interpolation using Algorithm **??**. However, different from roof edge images, hand written digit is much more complicated, manually marking control points is almost impossible. Due to complexity of the digit character, dozens of control points must be used to guarantee the quality of synthetic digital character. Under such high dimension space of control points, it makes it hopeless and inefficient to optimize control points using approach such as Algorithm **??**. Therefore, I adopt a different strategy and using an extra step to help constructing the correspondence between control points of two digital images.

To enable a interpolation with a better quality, a procedure called control point migration is adopted in my work to pair up control points of different images. The idea of the approach is to first compute an root image for all images with the same digit character which summarizes the digit character in all images. Then on this root image I could set up some control points which later could be migrated to all other images. Since for all images control points are from the same root image, their are already paired up.

A method called congealing proposed in [**?**] is adopted here to generate root images. In congealing, the project transformations are applied to images to minimize a joint entropy.

Thus the root image actually can be considered as an average image of all images after congealing. I show all the root images in Figure **??**.

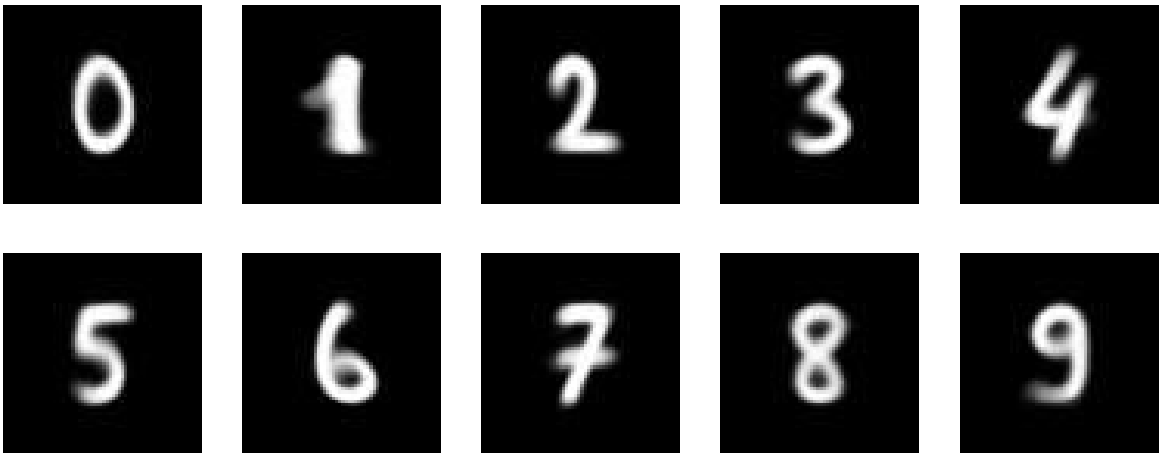


Figure 2.7. Root images of all kinds of digit characters.

Then control points are evenly sampled from the boundary detected from the prototype image. The control points needs to be mapped to each digit image in order to build correspondence with other images. To find this mapping I implement an approach that migrates the control points from the prototype images to destination image. Basically, given two sets of control points on two images, finding a matching between each pair of points is a bipartite matching problem, which may need some sort of features to be computed for each control points and could be solving using discrete optimization method. Additional constraint also has to be added to the solution such as matched point pairs has to keep their original order. Such constraints make the problem even harder to solve. In my method, however, the idea is very simple yet very effective that I slightly change source image a little bit towards the target image, and move the control points in the same time. my method achieve very good quality when intermediate steps are dense enough.

This point migration algorithm is based on a series of intermediate images generated in between synthetic prototype and destination image. To generate the intermediate images, I binarize all the images and the distance transformed images[**?**] of the synthetic prototype and the real image are generated. Given the number of steps, an intermediate image then

Figure 2.8. Illustrations of the migration of control points and intermediate synthetic images generated using control points in each step. The distance transform images of the synthetic prototype and real images are shown as the left most and right most images respectively.

is generated as a binarized image of linear interpolation between two distance transformed images. In each step, the control points are snapped to the closest boundary pixels of the intermediate image. The algorithm of OptimizeControlPoints($U, V, \mathbf{S}$) in this situation is given in Algorithm **??**, I fix the number of steps to 10 in this algorithm. The procedure of control points migration is shown in Figure **??** and Figure **??**. I recently notice that method proposed in [**?**] is very similar to my migration approach.

Algorithm 5 OptimizeControlPoints($U, V, \mathbf{S}$) Case 2

Input:

- A real image U.
- A prototype of the synthetic image $\mathbf{S} = \{\mathbf{P}, \mathbf{E}\}$.
- A synthetic image V.

1: $steps = 10$.
2: Compute distance transform image of U, V as U', V'.
3: **for** $i = 1$ to $steps$ **do**
4: $\quad I = (1 - \frac{i}{steps})U' + \frac{i}{steps}V'$.
5: $\quad I$=Binarize(I).
6: $\quad$ Update **S** by snapping to the closest boundary pixel on I.
7: **end for**
8: Set the status of **S** to be converged.
9: **return S**.

Once all images get their control points migrated from root image. More synthetic images could be generated using Algorithm **??**.

2.5 Creating Synthetic Data to Avoid Manual Labelling

One of requirement for a successful machine learning algorithm is abundant labelled data with high quality. Such dataset usually requires a long time and a lot of labor

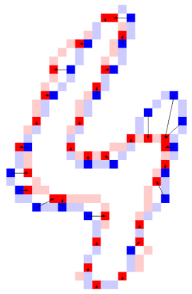


Figure 2.9. Illustrations of the migration of control points for character 4. The control points are migrated from root image (blue) to destination image (red) and arrows are used to indicate points moving direction.

work to be done, which make the labelling work very expensive. In some cases, it is even impossible to finish labelling such as the problem of point cloud roof style recognition I solved in [**?**].

in [**?**], I proposed a learning based roof style classification algorithm using aerial LiDAR point clouds. The proposed approach is able to classify complex roof styles which may be composed of simple roof styles including: curve, flat, gable, hex, hip, mansard, pyramid, shed, gambrel, dome and unknown. All roof styles classified in my work are listed in

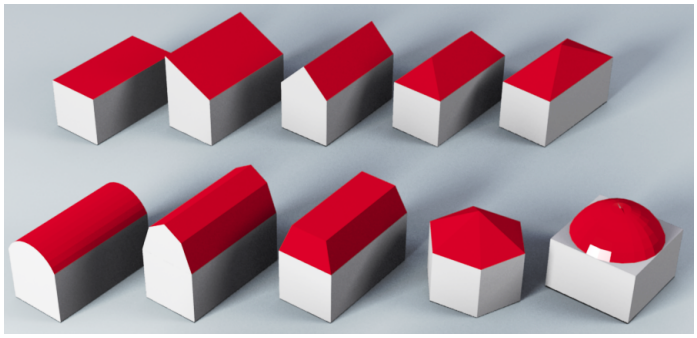


Figure 2.10. Roof styles I classified in [**?**]. From top to down, left to right: flat, shed, gable, hip, pyramid, curve, gambrel, mansard, hex and dome.

This method can easily accommodates more new roof styles by re-training of the new dataset with the new roof styles added. Because my method is implemented based on

a bag of works schema.

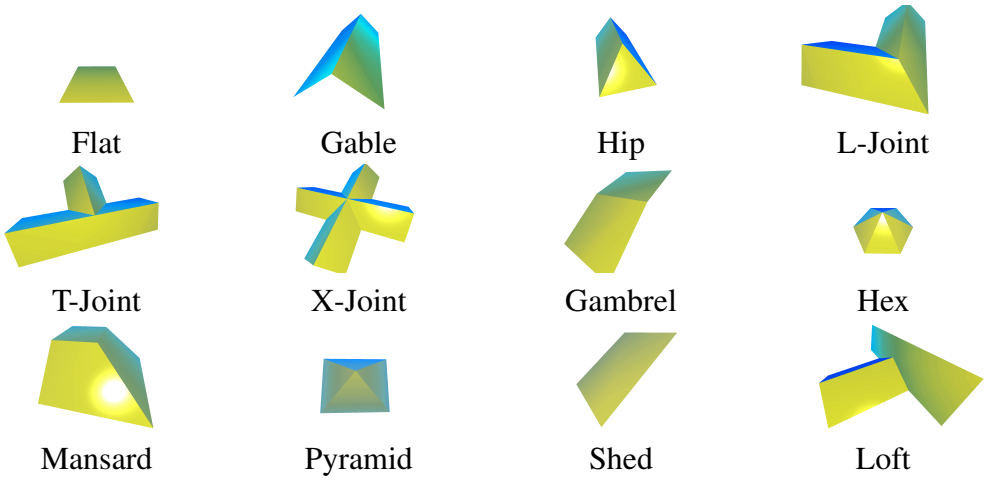


Figure 2.11. Illustration of all roof base models that are used in this work.

The difficulty in this work is to produce a codebook of important roof points. This codebook can then be used for a bag of words recognition. Learning these codebook using unsupervised learning is most straight forward. However, unsupervised learning is often misguided easily by the data and detects uninteresting patterns within the data.

Instead, I propose to integrate existing knowledge of roof structure and cluster the points of target roof styles into several semantic classes which can then be used as code words in the bag of words model. I use synthetic variants of these code words to train a semantics point classifier. Thus, I manually represent the codes in the codebook as semantic parts of the roof structure produced by manually analysing roof style models. I then learn a point semantics classifier using a large amount of synthetic variants of roof styles. There are 33 codes in the codebook used in this work. The key points that are used to produce the code words are shown in Figure **??**.

As can be observed in Figure **??**, some roof styles shown in Figure **??** are used to produce multiple key points. The synthetic variants of the codes are derived from these based models in two ways. First, I generate roof models by changing the parameters of the roofs including height, width and slope. Since it is inaccurate and unreasonable to assume

any distribution for these parameters and I want to generate models that can cover as much as possible cases in real world, I evenly draw these parameters from their given ranges. I present the size derivation of each base roof in below:

Flat: A little slop is allowed in my flat roof, so I change the angle between flat surface and ground plane from 0 degree to 10 degree, as an exception, distribution of roofs in between these two angles are even.

Gable: The opening width of gable roof is set to be 10 meters, I change two parameters for gable. Firstly, the median ridge is shifted, starting from center location and end at 2.5 meters from center. Secondly, by fixing opening width, the dihedral angle between slop surface and ground is changed from 20 degree to 75 degree, where mean is 30 degree.

Hip: In addition to how I derive from gable roof, the angle between side slop surface and ground plane is changed from 20 degree to 75 degree where mean is 30 degree.

L-Joint: For each of two branches, derivation of gable is applied.

T-Joint: For each of two branches(vertical and horizontal), derivation of gable is applied.

X-Joint: For each of two branches which cross with each other, derivation of gable is applied.

Gambrel: For two slop surface, the angle between surface tangent direction and z direction is changed. Suppose the angle of upper surface with z is a_0 and angle of lower surface with z is a_1. Using 45 degree as mean, a_0 and a_1 are changed from 20 to 75 degree, subject to $a_0 - a_1 \geq 20$ degree.

Hex: Synchronously change the slop angle of side face from 20 degree to 75 degree, where mean is 30 degree.

Mansard: Synchronously change the slop angle of two side face from 20 degree to 75 degree using 30 degree as mean.

Pyramid: Similar to Hex

Shed: Change slop angle from 20 degree to 75 degree using 30 degree as mean.

Loft: For slop surface, Change slop angle from 20 degree to 75 degree using 30 degree as

mean and for gable part, please refer to derivation of gable.

In addition to size variation, I also consider erosions of roofs, because it is common in my data to have eroded roofs due to unknown reasons. Four versions with different magnitude of erosion are derived using approach described in [**?**] for each generated roof which are shown in Figure.

Figure 2.12. Illustration of all semantic points used in my work, red sphere on each roof represent the typical location of semantic point, the color bar under each image correspond to all color labels used in previous demo.

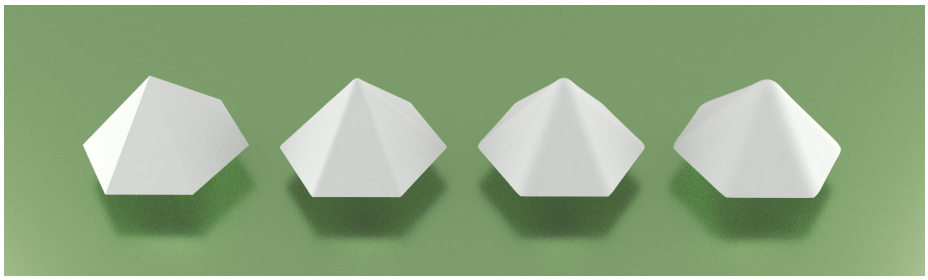


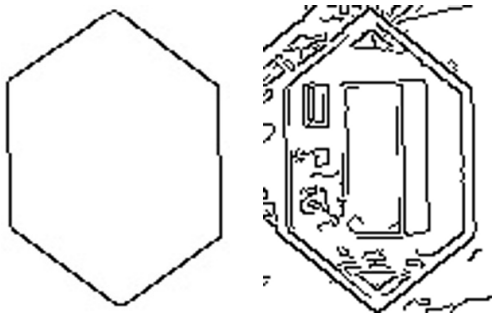
Figure 2.13. Illustration of erosion in my data. Three different levels of filtering are applied to each roof. In total, including original version, four versions are used in my work.

CHAPTER 3

LEARNING FROM SYNTHETIC DATA

3.1 Introduction

Directly use and learn synthetic data will most likely cause a failure or performance going downward in machine learning even the amount of synthetic data is far enough. Because, synthetic data is generated using parametric model which is a simplification of actual data. Thus, intrinsic patterns and noises owned by actual data do not appear in synthetic data. An example of actual roof edge image and corresponding synthetic roof edge image in my work [**?**] is shown in Figure **??**. It could be seen from the figure that synthetic building roof image characterize the primary structure only which is a boundary of hexagon. However, the actual roof image in real world comes with more edges extract from inside and edges extracted from other buildings outside the boundary. Some times, such differences could be caused by method of data collection itself, such as a the dataset I used in my work [**?**], that it is very hard for synthetic data to compensate these difference which is not necessary either. Actually, it is almost impossible to simulate every possible noise or inherent patterns in synthetic data using parametric model, which is not the intention of creation of parametric models in my method.



Actual Synthetic

Figure 3.1. An example of actual roof edge image and corresponding synthetic roof edge image in my work [**?**] is shown in above figures.

Such differences between actual data and synthetic data is very hard to eliminate in

data space and will cause a shift of synthetic data domain from actual data domain. Thus any machine learning framework learned from these synthetic data will most likely failed to work for actual data. Such shift is defined as a *Synthetic Gap* in my work [**?**] which will be discussed in detail in Chapter **??**. It is necessary for a learning procedure to remove or at least weaken such discrepancy in order to have a successful recognition result later on.

Information contained by actual data and synthetic data is not equivalent. Actual data could either contains more information in terms of noises and inherent patterns, or contain less information due to data loss such as occlusion in images. Synthetic data used in my work are created based on several parametric prototypes. The prototypes are composed of and simplified to contain the most basic and essential structure of actual data.

A general goal of feature extraction is to minimize the difference between actual data and synthetic data in extracted features. There two ways to achieve such goal First, extracted features has to be able to catch the most essential characteristics between actual data and synthetic data, thus ignore all the noises and inherent patterns contained . Second, extracted features should be able to compensate these additional information for synthetic data. Under these two directions, several techniques and algorithms are invented in my works.

3.2 Related Work

Most of existing works create synthetic data as a perturbation of original data [**?**][**?**][**?**][**?**][**?**][**?**]. Synthetic data under this domain is generated by either deforming original data using geometric transformation such as scaling, rotation, slant, shrinking ans so on or degradation model such as adding noise, erosion, removing parts and so on. Such data synthesis is very easy to implement in practise although, results in very limit data coverage in data space. Thus, generated synthetic data still carry the same amount of information and stay in the same disjunct as original data.

So far as I know, just very a few existing works treat synthetic data specially either in feature extraction or learning procedure. One of earliest works dealing with this problem is [**?**]. In this work, authors pioneer to use synthetic data in character recognition. They assume characters undergo a series of geometric transformation defined by them. To simulate real data, their recognizer is trained using synthetic data using their transformation. An interesting setup in their procedure is they apply a set of predefined inverse perturbations to the input image and are expected to include the true perturbation that actually made the input image different from its standard pattern. The corresponding inverse image will be very close the original standard pattern and could be easily recognized by some known method if an inverse perturbation actually corresponds to the true perturbation. Therefore in their learning pipeline, each inverse image is submitted separately to a conventional recognition system, the output score of which is then compared to others. It is clear that among the scores, the one corresponding to the true perturbation can be expected best. Since each score is attached to a class, the recognition scheme is in fact a by-product of the reversing process. A demonstration of this system is show in Figure **??**. Similar ideas are implemented in [**?**] and [**?**] in which degradation model or transformation model are trained and tested separately.

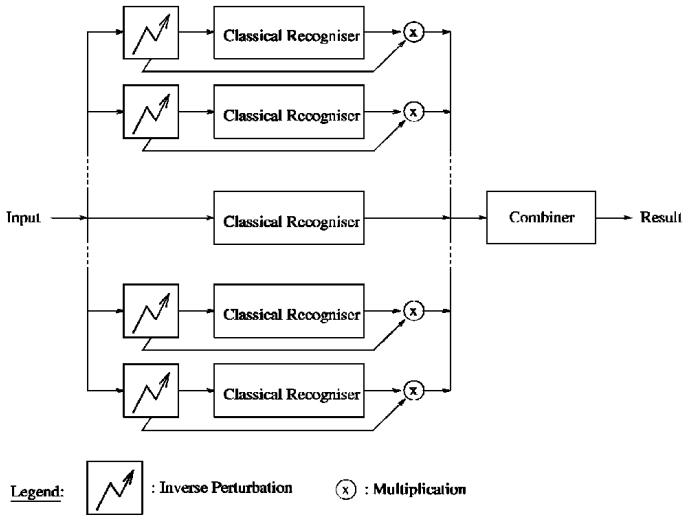


Figure 3.2. The schema of perturbation-based recognition of [**?**] is shown above.

Another strategy to handle the inherent difference between synthetic data and actual data is to use highly abstract features which to some extent could ignore divergence between two data and are able to extract the most common and essential characteristics from the two data. I notice such idea is applied in [**?**] where to build a optical character recognition (OCR) system, a Kolmogorov complexity distance is used to measure the similarity between two data. Kolmogorov complexity measures the amount of information contained by each data which is done by computing the highest compression ratio of data when being compressed. By using this feature, feature extraction process ignore the concrete differences between data and focus on most essential information contained in the data.

Most of existing works generate synthetic data in feature space. Most of techniques are invented to generate and learn synthetic data in feature space. I put the discussion of using synthetic data in feature space to Chapter **??** and use that chapter to present more details of algorithms and strategies.

3.3 Features to Ignore Additional Information in Actual Data

3.3.1 Roof Style Classification. In my work [**?**], my goal is to classify actual satellite roof edge images using synthetic edge images I created in Chapter **??**. The synthetic data is created as connected line segments on image to represent roof main structures such as ridge and valleys. In actual roof edge images, due to low image quality in original source images and noises coming from for example shadows, occlusions and small objects on roofs, the edges representing main roof structures are composed of intermittent short segments and a lot of random size circles. Due to these defect, traditional image features such as Histogram of Gradient (HOG), Local Binary Pattern (LBP) do not work very well on minimizing the difference between actual data and synthetic data.

To facilitate a better learning from synthetic satellite roof edge images, I propose a

Table 3.1. Precision and recall of hip (HIP), gable (GBL), flat (FLT) and half hip (HHP) styles classified by Random Forest is shown in above table. I use red font to show highest value among results.

	Real Roof		Combination	
	Precision	Recall	Precision	Recall
HIP	0.953	0.823	0.963	0.814
GBL	0.877	0.901	0.886	0.882
FLT	0.784	0.977	0.745	0.959
HHP	0.893	0.439	0.967	0.509

new image feature called Histogram of Oriented Rays (HOR). HOR works by tracing the longest continuous ray in each direction centerred at each pixel. A few parameters are set to enable HOR to be tolerant to small gaps along the ray and a small range of perturbation of the pixels on the ray. Similar to HOG feature, HOR uses a histogram to save information from all orientations around a pixel, shown in Figure **??**. The exact algorithm of HOR is given in Algorithm

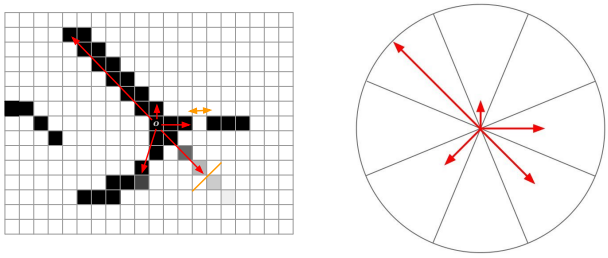


Figure 3.3. Illustration of HOR feature.

To evaluate the performance of HOR, a test set containing 1783 roof images including flat, hip, gable and half hip roof styles is tested. I use random forest as a base learner and compare results to features including HOG, Shape Context (SC), LBP and so on. A same feature merging rule is used in all tests that each feature descriptor has 5 pixels per cell, 2 cells per block and 50% overlapping between blocks. The precision and recall obtained using HOR feature are show in Table **??**.

I show the accuracies of each feature using my testing data classified by the classifier, which trained by the combination of my training dataset and synthetic models. The

Algorithm 6 Histogram of Rays

Input:

Given point p on gray scale image I;
Number of search orientation D;
Searching threshold T;
Mimimum gap length G;
Length of HOR feature vector L;

Output:

The HOR feature vector V of point I;

for $i = 1$ **to** D **do**
  Terminator $count = 0$;
  Calculate pixel list on direction $\frac{D}{2\pi} \Rightarrow IDX$ from p;
  Index $j = 0$;
  while $count < G$ **do**
    if $IDX_{j+1}/IDX_j < T$ **then**
      $count++$;
    else
      $count = 0$;
    end if
    $j++$;
  end while
  $V_i = j$;
end for
Sampling V from length D to L;
return V;

accuracies is presented in Table **??**. The precisions calculated by using HOR and HOG combination for hip, gable, flat and half hip(93.1%, 95.5%, 98.2% and 66.7%) shows a better result than using HOR(89.8%, 95.0%, 96.8% and 63.2%) and HOG(80.5%, 88.2%, 95.4% and 59.7%) separately. The results show that for current edge detection approach and training method, HOR alone, or HOR combined with HOG, are considered to be more appropriate as the input feature for roof style recognition system.

There is no roof labelled as half hip in the subset weevaluate of my sponsors dataset. The only one I obtain is in Figure 1, where captured in Stanford campus. I cant say the accuracy for half hip style is 100% based on this ideal same. The final accuracy of roof in complex footprint recognition is 95.2%, 84.6% and 91.7% for flat, hip and gable styles.

Table 3.2. The comparison of accuracies between HoG (HOG), Shape Context (SC), HoR (HOR) and their combinations by running Random Forest approach.

	HIP	GABLE	FLAT	HALFHIP
HOG	0.805	0.882	0.954	0.597
SC	0.350	0.828	0.959	0.140
HOR	0.898	0.950	0.968	0.632
LBP	0.000	0.986	0.631	0.000
HOR+HOG	0.931	0.959	0.982	0.667
HOR+SC	0.619	0.891	0.959	0.436
HOG+SC	0.752	0.959	0.945	0.474
SC	0.743	0.869	0.963	0.509

3.3.2 Satellite Building Image Retrieval. An extension of classic chamfer matching algorithm is proposed in my another work [**?**]. In this work, I deal with a image retrieval problem for edge images extracted from building roof top. Different from roof top images I used in other works, roof top images in [**?**] are extracted from artificial building footprint, thus, the edge images contain only boundaries of a roof. An example of target roof image and synthetic roof image is given in **??**.

The task in this work is to retrieve the most matching actual roof edge image given a synthetic roof edge image. Challenges come from different aspects. First, actual edge image are generated by extracting edges from satellite images, thus edge images could contain a great number of edges from other unrelated objects in image. In addition, due to lighting condition, buildings in image could either be too dim or shadowed by dark areas. Second, synthetic edge image basically are polygons representing the boundary of a building. These polygons give a very approximate shape about an actual building, thus, the polygon could be very different from the actual building roof top.

In order to be resistant to all kinds of defects in actual roof edge images, I extend the classic chamfer matching feature and design a new chamfer matching that is much more robust when facing an image retrieval problem such us the one in my work.

The idea of using Chamfer matching for image registration was first introduced by

Barrow *et al.* [**?**] where they try to find the model of a coastline in a segmented satellite image. Since then many variations of the Chamfer matching have been introduced.

An important variation of the Chamfer matching by Borgefors [**?**] uses a hierarchical Chamfer matching algorithm (HCMA). This algorithm uses an image pyramid in searching for the optimal position of a template. The search for an optimal position is made in different resolution levels of the pyramid by using a representation of the distance image. The optimization of the objective function is done by discretizing the transformation parameters and stepping through them in each pyramid level. The speed of HCMA can be improved by modifying the computation of the distance transform image and by selecting the starting search position [**?**]. An HCMA algorithm based on interesting points has been used in [**?**] where a parallel computation scheme of Chamfer matching is discussed. The selection of interesting points in this work is done through a dynamic threshold scheme guided by a histogram.

A Chamfer matching algorithm which is based on multiple features was introduced by Gavrila et al.[**?**],[**?**], where it is proposed to use edge orientation as a feature. An orientation channel is created for each feature and a distance transform image is generated for edges in the channel. This method uses a hierarchical scheme for matching multiple templates with an image in which similar templates are grouped at different levels.

Shotton et al.[**?**] use the Chamfer distance with an additional cost which measures the mismatch of edge orientations given by the average difference in orientation between template edges and the closest edges in target image. Instead of explicitly formulating a separate term of orientation mismatch, the orientation difference is generalized in the computation of the Chamfer distance[**?**]. A comparison between shape context matching and Chamfer matching is conducted in [**?**], where results show that using the Chamfer matching is faster than matching using shape context and that global matching using Chamfer matching is better than using shape context.

3.3.3 Chamfer Matching. Let U and V be two binary edge images, where U is the target image and V is the template image. Let $\{u_j\}_{j=1}^{m}$ and $\{v_i\}_{i=1}^{n}$ be the edge pixels in these images respectively. Let $U(u_j)$ denote the value of image U at location u_j.

In chamfer matching, I seek a correspondence between $\{u_j\}$ and $\{v_i\}$ under transformation W. Assuming that the transformation W between the template and target images is rigid with translation T and rotation R, a template edge pixel v_i is transformed into the target image by using following expression:

$$W(v_i;\ R,T) = R \cdot v_i + T \equiv v_i^U \tag{3.1}$$

Given a distance metric $d(\cdot)$, I can solve for the transformation parameters T and R by minimizing the total distance:

$$D(U,V;\ R,T) = \frac{1}{n} \sum_{1 \le i \le n} d(v_i^U, u_j) \tag{3.2}$$

where u_j is the closest target edge pixel to v_i^U in the sense of the distance metric $d(\cdot)$.

The computation of $d(v_i^U, u_j)$ can be done in linear time using the *distance transform* of the target image[**?**]. Denoting the distance transform image of U by $U_{\mathcal{DT}}$, I can write Equation (**??**) as:

$$D(U,V;\ R,T) = \frac{1}{n} \sum_{1 \le i \le n} U_{\mathcal{DT}}(v_i^U). \tag{3.3}$$

Various types of distance transforms can be produced using different distance metrics $d(\cdot)$. In [**?**], city-block distances are used and a two-step linear algorithm is proposed to compute the distances. Euclidean distances approximations are provided by Borgefors[**?**], Montanari[**?**], and Danielsson[**?**]. An efficient squared Euclidean distance computation al-

gorithm is described by Felzenszwalb[**?**], Felzenszwalb's algorithm can be generalized to compute other distances.

3.3.4 Extended Chamfer Matching. I extend the Chamfer matching in several ways. First, I extend the basic Chamfer matching and obtain additional robustness by jointly minimizing the spatial distance between pixels and the misalignment of edge orientations. A similar distance metric is used in [**?**] where a linear representation of edges is generated to model the edge orientation. The proposed approach avoids the need for a linear representation of edges by computing the edge orientation directly from the images.

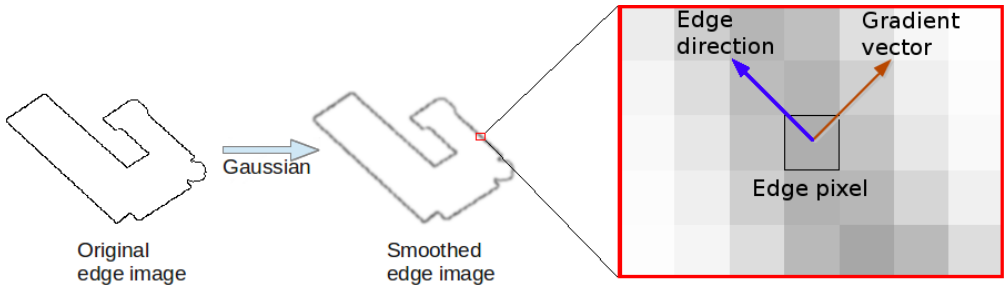
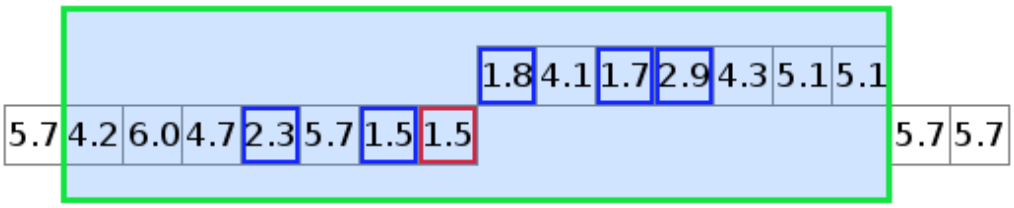


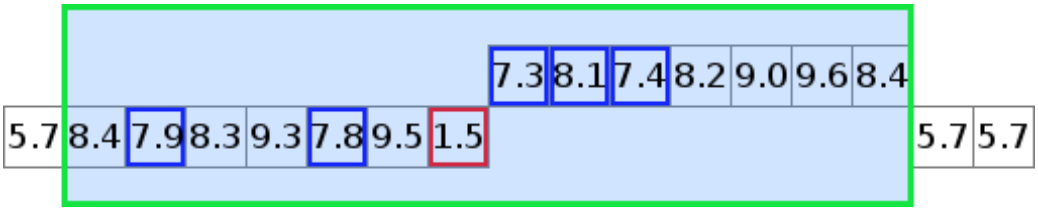
Figure 3.4. Computing edge orientation in the binary template image.

The edge orientation at each location is computed as a vector perpendicular to the gradient vector at that location. In the target image, the distance transform image $U_{\mathcal{DT}}$ is used to compute the edge orientation at each location. Specifically, given a transformed location v_i^U, its edge orientation $\vec{v}_i^U$ is computed as a vector perpendicular to the gradient vector $\nabla U_{\mathcal{DT}}(v_i^U)$. To compute the edge orientation in the binary template image V, I create a gradient vector field around edges by smoothing V using a standard Gaussian kernel. The edge orientation vectors $\vec{v}_i$ are then calculated as vectors perpendicular to the gradient vectors obtained from the gradient vector field. An example of the edge orientation computation in V is shown in Figure **??**.

Having the edge orientation computed in both the target and template images, the distance between pixels v_i^U and u_j is computed by:



(a) Modified distance = 1.5 × 0.55 = 0.82



(b) Modified distance = 1.5 × 2.54 = 3.82

Figure 3.5. Example of the modified distance measure. In both cases I use a neighborhood of size $p = 13$, and select the lowest $q = 5$ neighbors. While the distance at the pixel is the same (1.5), the modified distance measure in the bottom example is higher due to the larger distance to neighbors.

$$d(v_i^U, u_j) = \lambda U_{\mathcal{DT}}(v_i^U) + (1 - \lambda)(1 - \| \cos(\alpha_{v_i^U}) \|) \tag{3.4}$$

where λ is a weight factor that controls the importance of orientation mismatch, $\cos(\alpha_{v_i^U}) = \langle \vec{v}_i^U, \vec{v}_i \rangle$ measures the orientation mismatch, and it is assumed that $\vec{v}_i^U$ and $\vec{v}_i$ are normalized.

A squared Euclidean norm is used in the first term of Equation (**??**), which gives a larger penalty to mismatched pixels. A method to generate the squared Euclidean norm distance transform in linear time is described in [**?**].

3.3.5 Edge Distance Variance. To have a robust matching result, the matching error of a pixel v_i should not solely depend on $d(v_i^U, u_j)$. This is because it is possible that v_i^U will be an incorrect edge pixel in the target image that happened to have a small error. Based on this consideration, I argue that to get a confident measurement of the matching error it is necessary to take into account the matching error of neighbouring pixels.

Given the transformed template edge pixels $\{v_i^U\}_{i=1}^n$, I find for each pixel v_i^U matching scores for the p closest transformed template edge pixels. To estimate contextual matching error of $d(v_i^U, u_j)$ at location v_i^U, the q pixels with lowest matching error, where $q < p$, are selected. I denoted these q pixels as $\{v_k^U\}_{k=1}^q$. I then calculate the variance of the distance $d(v_i^U, u_j)$ of the q selected pixels:$\varphi(v_i^U) = \frac{1}{q}\sum_{k=1}^q (d(v_k^U, u_j) - \bar{d})^2$, where $\bar{d}$ is the average matching error of the q neighbors.

The distance variance $\varphi(v_i^U)$ provides a more stable assessment of the matching result at v_i. The parameters p and q control the size of the contextual information used in the computation. A larger p leads to more contextual information included while q helps in excluding outliers. In my experiments, I set $p = 13$ and $q = 5$. Since p and q are constants, the asymptotic time complexity of the algorithm is not affected by them. From a practical point of view, to save time when searching for the p closest neighbors at each v_i, I generate and maintain a list of p closest neighbors for each v_i before the matching begins.

As I would like to give preference to distance measures with small distance variance, I modify the distance metric in Equation (**??**) by multiplying it by a factor which is proportional to the distance variance:

$$d_\varphi(v_i^U, u_j) = d(v_i^U, u_j) \times (1 + \varphi(v_i^U)) \tag{3.5}$$

Figure **??** shows an example of the modified distance computation.

The significance of multiplying the distance variance in Equation (**??**) lies in several aspects. First, the distance measure is no longer solely dependent on each pixel's individual matching distance as the distance now considers the relationship between a pixel and its best matched neighbours. Consequently, I expect each pixel and its connected neighbours to have a small distance. Second, the variance in Equation (**??**) makes it much easier to

separate well matched pixels from mismatched ones by increasing the width of boundary that separates the well matched pixels and mismatched ones.

I tested the proposed approach on two datasets, where each contains 1000 building models selected from a San Francisco (SF) and a Chicago (CHI) urban area.

The four parameters in Algorithm **??** are set as follows: $\theta = 50\%$, $t_s = 5$, $t_\alpha = 15$, and $t_\varphi = 0.8$. By setting these parameters, I allow for at least 50% of template pixels to be inliers in the computation of $D(U, V;\ R, T)$; the accepted matching error of obtained transformation has to be within error of 5 pixels in terms of spatial distance and 15 degrees in terms of orientation difference; finally 0.8 average distance variance is used to rule out outliers during Chamfer matching.

In my experiments, I test the accuracy of my algorithm in estimating the optimal transformation during alignment. I manually labelled the ground truth locations of the building roofs in satellite images. The accuracy of the result is measured by the ratio of overlap between the bounding box of the ground truth roof mask and the one transformed by the proposed algorithm. Note that simply measuring the Root Mean Square error (RMS) between the aligned targets and templates is not accurate as it is sensitive to broken edges and incorrect alignments.

I first test the proposed Chamfer matching without using global constraint. I divide my evaluation into two parts. In the first part, the evaluation is done on all the buildings of both datasets using the same λ. I then change λ between 0 and 1 to compare the robustness of the proposed approach, and evaluate the relative importance of the spatial and angular terms. In the second part, to assess the performance of the proposed algorithm on matching occluded objects, I specifically run tests on building images in which the target buildings are partially occluded. In all the tests, I compare my results with that of the directional Chamfer matching (DCM) proposed in [**?**] and the basic Chamfer matching (CM). Note

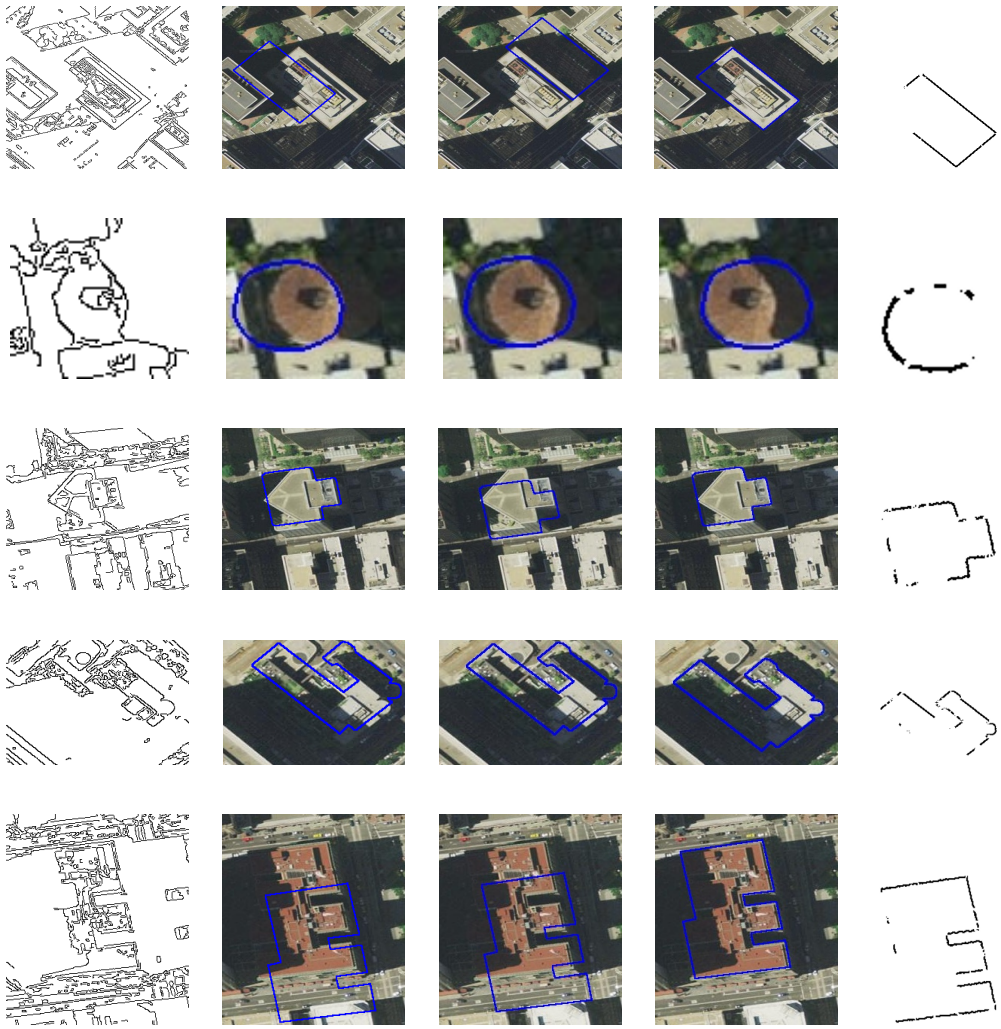
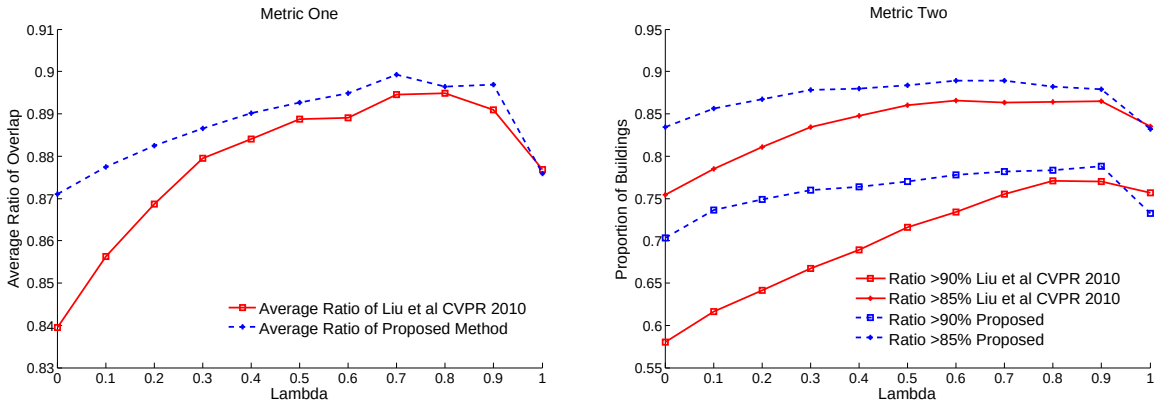
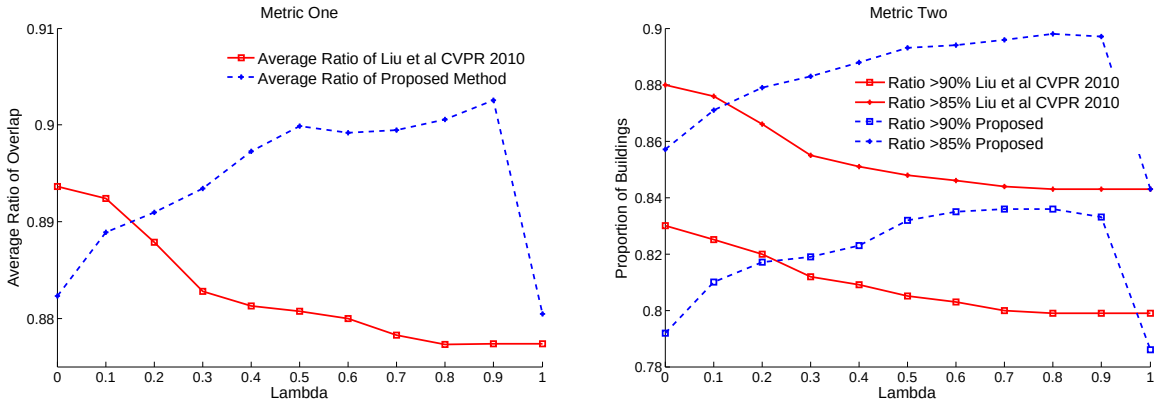


Figure 3.6. Results of matching partially occluded buildings. (a) Edge images detected from target image. (b)-(d) Results of the CM algorithm, the results of the DCM algorithm and the results of the proposed algorithm, respectively. (e) The pixels selected for computing the average error during the matching process.



San Francisco



Chicago

Figure 3.7. Experimental evaluation results on the San Francisco (upper row) and Chicago (lower row) datasets. The left and right columns show different evaluation metrics.

that the result of CM is a special case of the DCM algorithm when $\lambda = 1.0$.

Two kinds of metrics were used in my evaluation. In the first metric, I compute the average ratio of overlap area in the two datasets. In the second metric, I consider only results with area overlap above a certain rate (85% and 90%).

In the first test where all the buildings were included, it could be observe in Figure **??** that my algorithm generates better results compared with DCM and CM at almost every setting of λ. In both datasets, my algorithm maintains a similar and consistent performance, while DCM performs differently on the two datasets.

Table 3.3. Accuracy results.

		Metric One	Metric Two $\geq 85\%$	Metric Two $\geq 90\%$
SF	Proposed	91%	0.94	0.90
53 Buildings	DCM	63%	0.33	0.28
	CM	35%	0.20	0.18
CHI	Proposed	80%	0.81	0.78
74 Buildings	DCM	51%	0.21	0.20
	CM	32%	0.16	0.13

To test proposed algorithm's strength in finding matching targets under partial occlusion, I specifically test and evaluate the algorithms' performances on 53 buildings from San Francisco and 74 buildings from Chicago which are significantly occluded by shadows. I use a parameter of $\lambda = 0.7$ in this test. The results are shown in Table **??**. As can be observed, the proposed algorithm achieves 80% accuracy on both datasets compared with about 50% accuracy when using DCM and 30% accuracy when using the original Chamfer matching. Some examples of matching incomplete buildings are shown in Figure **??**.

I finally tested the improvement obtained by the global constraint for $\lambda = 0.7$.For both data sets, I obtain at least 2% improvement in terms of accuracy. The comparisons of alignment results with and without global constraint alignment are given in Table **??**.

Table 3.4. Results of alignment using global constraint. The number on the right side of the arrows show the results obtained by alignment using the global constraint.

	Metric One	Metric Two $\geq 85\%$	Metric Two $\geq 90\%$
SF	$89\% \rightarrow 91\%$	$88\% \rightarrow 90\%$	$78\% \rightarrow 79\%$
CHI	$89\% \rightarrow 92\%$	$89\% \rightarrow 91\%$	$83\% \rightarrow 85\%$

3.4 Features to Compensate Additional Information

There are cases when ignoring additional patterns in actual data are very difficult. An typical example is given in my work [**?**] where the exact structure of actual data is hard to be recognized thus there is no way under this scenario to find corresponding geometry of synthetic data. Therefore, rather than designing feature to neglect additional patterns on the side of actual data, I am going to learn additional patterns on actual data and design features which are able to encode these information in the computation.

This done from different aspects in [**?**]. First, features in this work are designed to characterize spatial and context information of local point.

Spatial features take into account the neighborhood of each point. So the spatial features are powerful to describe geometry characteristics in the point's neighborhood and so assist in producing a distinct characterization of it. Let p_i be a key point. Let N_i be a neighborhood of the point, containing neighbors in a radius of μ (0.6 meters in my experiments). I denote the spatial features at p_i using $\mathcal{S}_i$.

i. Eigen Features (EF). I compute the eigenvalues: $\lambda_1, \lambda_2, \lambda_3$ of the covariance matrix of neighbors N_i centred at p_i. I then add following features for p_i: λ_1, λ_2, λ_3, $\lambda_3 - \lambda_2$, $\lambda_2 - \lambda_1$, $\lambda_1/(\lambda_1+\lambda_2+\lambda_3)$, $\lambda_2/(\lambda_1+\lambda_2+\lambda_3)$, $\lambda_3/(\lambda_1+\lambda_2+\lambda_3)$, yielding 8 features in total. These features capture the non-planarity of points around p_i.

ii. Point Feature Histogram (PFH). These are point cloud features based on histogram which are described in [**?**]. Taking into account computational efficiency, 3 subdivisions of

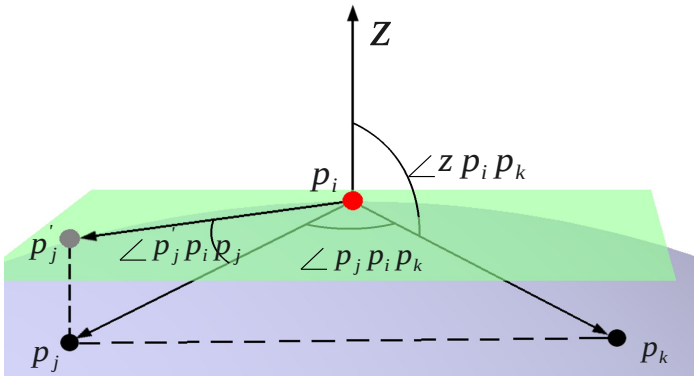


Figure 3.8. Illustration of the frame I used in the computation of the shape distribution features.

the features range are used in the feature histogram which yield $3^3 = 27$ features.

iii. Shape Distribution Features (SD). Shape distribution [**?**] measures global geometric properties of an object by representing object features as a probability distribution. Shape distribution is invariant to translation, rotation and scale and is highly informative in matching objects. Using the ideas of shape distribution I construct four features:

1) A2: Measures the angle between two vectors composed by two neighbors $p_j, p_k \in N_i$ and p_i, shown as $\angle p_j p_i p_k$ in Figure **??**. A2 feature is computed for all pairs of points chosen from N_i.

2) Az: Measures the angle between the z direction and a vector pointing from p_i to one neighbor p_k, shown as $\angle z p_i p_k$ in Figure **??**. All points in N_i are used to compute this feature.

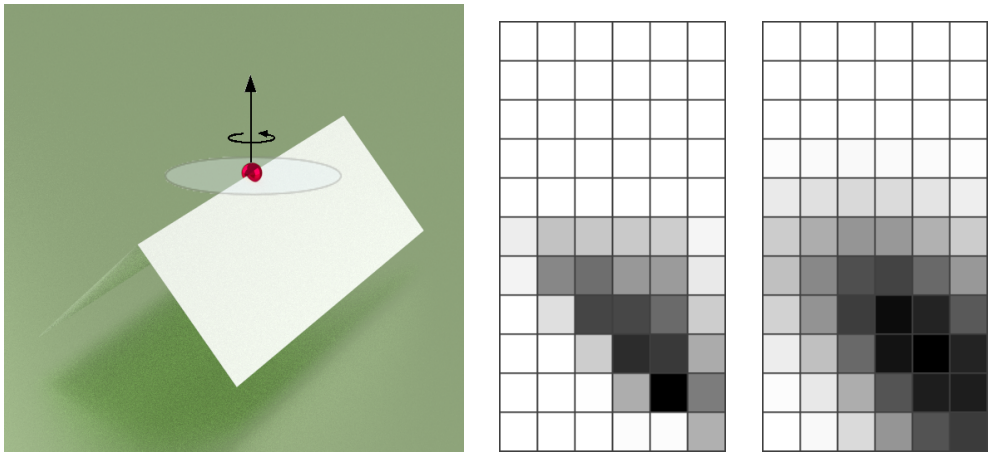
3) D2: The is the feature D2 as described in [**?**]. It measures the distance between any two neighbors of p_i, an example is shown as $\|p_j p_k\|$ in Figure **??**. D2 is computed for all pairs of points chosen from N_i.

4) Dt: Measures the angle between p_i's tangent plane and a vector pointing from p_i to another neighbor p_j, shown as $\angle p_j' p_i p_j$ in Figure **??**, where p_j' is the projection of p_j on

tangent plane. All points in N_i are used to compute this feature.

I use a histogram with 10 bins to represent each of the features described above. In total there are 40 SD features that are computed.

iv. Spin Image. Spinning around the z direction, I compute a spin image [**?**] with $6(width)\times 11(height)$ dimensions. Totally 66 features are contributed by spin image.



(a) (b) (c)

Figure 3.9. Example of applying Gaussian smoothing to the spin image features. (a) The red dot shows the location where spin image features are computed. (b) Generated spin image without smoothing. (c) Result of spin image after random disruption.

Synthetic data are created as locally perfectly planar or intersection of perfect flat surfaces. However, this usually is not a case in actual data where surfaces are locally irregularly bumpy and disrupted by outliers and noises. To better encode these irregularities in the spatial features, I learn these bumpiness from actual data and simulate them in synthetic data when computing above spatial feature.

In my work bumpiness is defined as the distance a point moving up and down along vertical direction. To model the bumpiness in my work, I make following assumptions. First, bumpiness is correlated to slope of local surface where points on flat surface are less likely to have bumpy effect than points on the slope. Second, bumpiness follows a normal distribution on direction of point's normal for points on the similar slope surface, shown in

Figure **??**. These assumptions though are quite heuristics, it make sense to have them in this work. Because laser signal decay much faster when bouncing back from slope surface thus the noise and outlier variance is larger for points on a surface with larger slope.

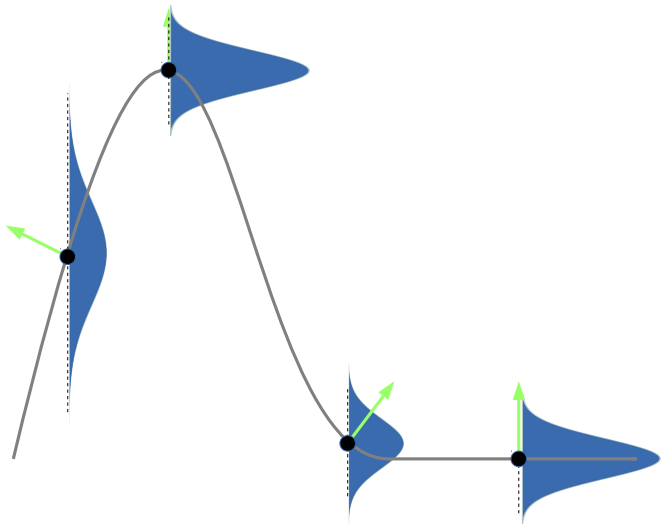


Figure 3.10. Illustration of distribution of bumpiness at points with different slope.

To model bumpiness, I assume entire data collection process is a stochastic process in which I assume for points with similar slope, point offset caused by noise is a Gaussian noise with zero mean $\Xi = \{\xi_i\}_{i=0}^{n}, \Xi \sim N(0, \sigma_i)$. For each point p_i, the noise associated with the point p_i is then computed as vertical distances to plane surfaces fitted by each of its neighbors. Denoting local plane surface fitted at point $p_j, p_j \in N_i$ as $F(p_j)$, where N_i is set of neighbors of p_i. Then the offset of p_i can be computed as a expectation of vertical distance to planes fitted by all its neighbors:

$$\text{Exp}(D(p_i)) = \sum_{p_j \in N_i} D_{F(p_j)}(p_i)\Phi(\|p_i - p_j\|) \tag{3.6}$$

where $\Phi(\cdot)$ is a weighting function that $\sum_{p_j \in N_i} \Phi(\|p_i - p_j\|) = 1$, and $D_{F(p_j)}(p_i)$ represents the vertical distance from p_i to a plane fitted locally at p_j.

To model the relationship between point offset and its slope, the slope of local

surface is computed as the dihedral angle between point normal and horizon. Given the maximum and minimum of the slope of all points, a histogram $H_s = \{h_i\}_{min(\text{Exp}(D(p_i)))}^{max(\text{Exp}(D(p_i)))}$ is built, Suppose k is the number of bins in the H_s and I use $H_s(i), 1 \leq i \leq k$ to denote the centring value of point slopes in each bin. Later, H_s is voted using $\text{Exp}(D(p_i))$. Then variance $\text{var}(h_s(i))$ is computed for distances within each bin $h_s(i)$. Therefore, I could have k tuples of $(H_s i, \text{var}(h_s(i)))_{i=1}^{k}$ for which I assume a polynomial relationship exists. Thus, a polynomial modal can be fitted to the tuples. An example of polynomial model with degree 2 is shown in Figure **??**.

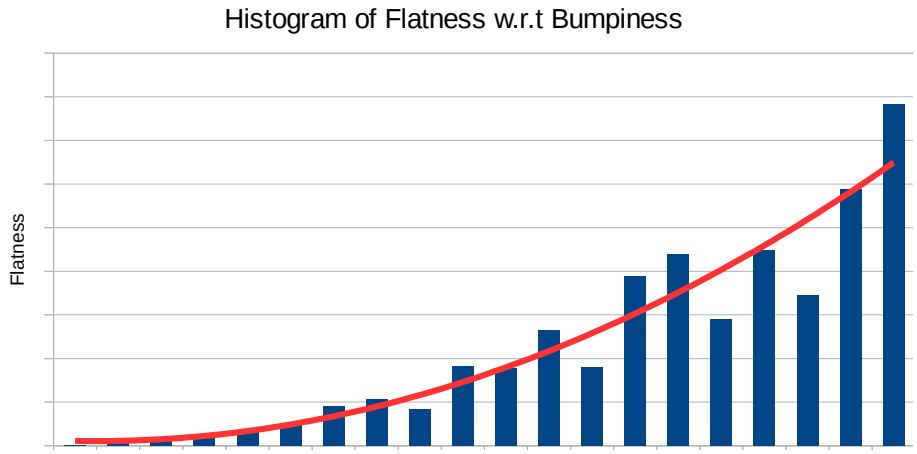


Figure 3.11. Illustration of the histogram H and the quadratic function fitted on it.

To this end, I have obtain the polynomial relationship between point slope and the variance of distribution of point offset with which I can encode this randomness in features generated from synthetic data. Given a synthetic roof, I first compute slope of every point on the roof, after which a adjustment of points location along vertical axis is applied to each point using polynomial model just obtained. Such position adjustments have been applied to all features introduced previously. An example in Figure **??** shows a comparison of spin image between using point adjustment and without.

The classification of roof style in this work has two stages. In the first stage, I classify points according to their semantics. A random forest is used in the first stage. The

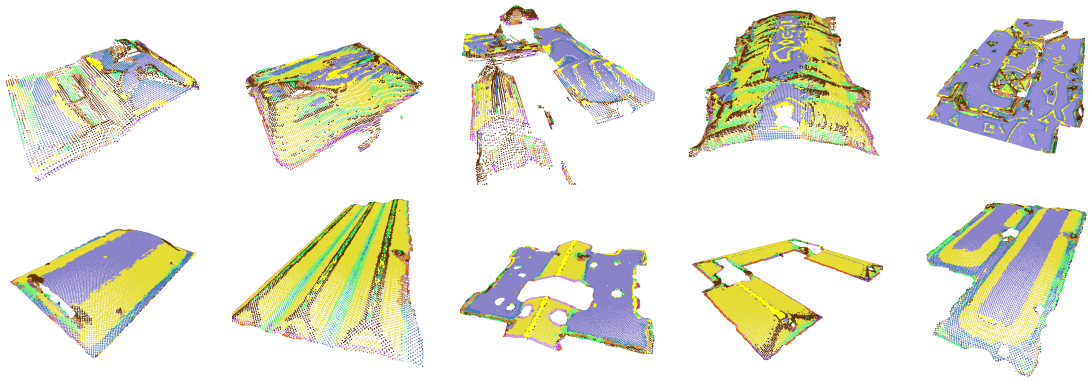


Figure 3.12. Point classification results obtained by running my point type classifier on dataset one (top row) and the dataset two (bottom row). The colors of points corresponds to the color bars of the point in Figure **??**.

classifier produces a probability measure for each of the 33 code words. Thus given a point p_i I have $\{P(l_j|\mathcal{S}_i)\}_{j=1}^{33}$, where $\mathcal{S}_i$ is the spatial feature set at p_i and l_j is the j-th label $j \in [1, 33]$. There are two parameters in setting the random forest: the number of trees ρ_0 in the forest and the number of features ρ_1 that the random forest can choose at each node. Assuming that the dimension of the input feature vector is ϵ, so I set $\rho_0 = 100$ and $\rho_1 = \sqrt{\epsilon}$. A histogram $\mathcal{H} = \{h_j\}_{j=1}^{33}$ is used to count the frequency of the points in a roof. Given points of a roof $\mathcal{P} = \{p_i\}_{i=1}^{n}$, each bin h_j of $\mathcal{H}$ is computed as:

$$h_j = \frac{1}{n}\sum_{i=1}^{n} P(l_j \mid \mathcal{S}_i, \mathcal{C}_i), \quad 1 \leq j \leq 33 \tag{3.7}$$

An example of the classification of roof points by the proposed approach is shown in Figure **??**. As can be observed in this figure, context features produce more regular and confident result. After convergence, the histogram $\mathcal{H}$ is used as the bag of words features of the roof.

The roof style classifier uses the bag of words features $\mathcal{H}$. I use a random forest classifier to classify each roof style into one of 9 possible roof styles. To accommodate various kinds of point cloud degradations in different datasets, the roof style classifier is

trained using a real roof dataset. I set the two parameters of the random forest as: $\rho_0 = 100$ and $\rho_1 = 6$. To avoid bias towards a particular roof style and create a balanced number of training data among different styles of the roofs, I super sample the training data using the SMOTE [**?**] algorithm.

The proposed approach has been tested on two datasets. In the first dataset, there are 3290 buildings that were extracted from Chicago urban area. In the second dataset there are 3290 buildings that were extracted from a San Francisco urban area. The two datasets have different characteristics and kinds of degradations.

In the first dataset, roof points are irregular and the shape of the roof is decayed to some extent. This is due to the fact that roofs in this dataset were originally produced from a highly down-sampled aerial LiDAR. The roof points in the second dataset have relatively low resolution and uneven distribution across the roof surface. Both datasets contain roof points of building only. The roofs were labelled to one of 9 target roof styles according to their appearances. A roof is labelled as UNKNOWN when either the roof is not recognizable or its roof style can not be categorized into one of the 8 styles. When a roof is composed of multiple styles, I label the roof according to the style of the largest component. The distribution of roof styles in the two datasets is shown in Figure **??**.

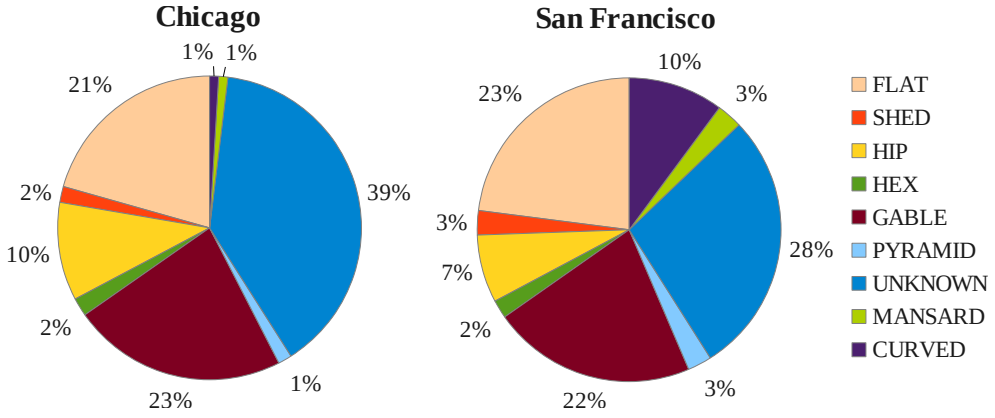


Figure 3.13. The distribution of the roof styles in the two datasets.

To evaluate the performance of my approach and measure the improvement in using

Table 3.5. Precision and recall of Gaussian Mixture Model(GMM), K-Means(KM) and my approach are shown in above table. I use red font to show highest value among results obtained by these three approaches.

	Chicago						San Francisco					
	——-Precision——-			——-Recall——-			——-Precision——-			——-Recall——-		
	GMM	KM	Ours	GMM	KM	Ours	GMM	KM	Ours	GMM	KM	Ours
FLAT	0.87	0.85	0.92	0.88	0.90	0.90	0.88	0.88	0.86	0.93	0.93	0.93
SHED	0.94	0.94	0.91	0.57	0.64	0.78	0.87	0.91	1.00	0.43	0.68	0.68
GABLE	0.62	0.67	0.71	0.86	0.84	0.88	0.57	0.61	0.65	0.71	0.69	0.77
HIP	0.65	0.63	0.63	0.16	0.36	0.37	0.55	0.61	0.70	0.22	0.28	0.31
HEX	0.87	0.86	0.93	0.87	0.81	0.90	0.90	0.80	0.92	0.83	0.66	1.00
PYRAMID	0.83	0.66	1.00	0.20	0.16	0.37	0.87	0.83	1.00	0.87	0.93	1.00
MANSARD	1.00	0.75	1.00	0.25	0.18	0.31	0.50	0.66	1.00	0.05	0.11	0.41
CURVED	1.00	0.93	1.00	0.87	0.93	1.00	0.71	0.70	0.74	0.77	0.71	0.79
UNKNOWN	0.84	0.85	0.97	0.88	0.85	0.90	0.62	0.59	0.66	0.63	0.64	0.66
Average	0.85	0.79	0.89	0.62	0.63	0.71	0.72	0.73	0.84	0.60	0.63	0.73

the synthetic model when generating the codebook, I compare the roof style classification results obtained by my approach to the results obtained by the K-Means (KM) and the Gaussian Mixture Model (GMM) algorithms for generating the codebook.

I evenly divide the dataset into two parts, one part is used as a training set, while the second part is used as a testing set. In the case of the KM and GMM algorithms, the codebook is generated using the training set. For each approach, the bag of words features of all buildings are then computed using its own codebook. I set $k = 30$ in the K-Means algorithm and $k = 27$ in the Gaussian Mixture Model algorithm as these were the two configurations that provided the best accuracy for these approaches. The roof style classifier of each approach is then trained and tested using its own bag of words features. The results of these three approaches in terms of precision and recall are shown in Table **??**. F-Score results are shown in Figure **??**.

As can be observed, the proposed approach performs better for almost all roof styles. Considering the roof styles of PYRAMID and MANSARD, I observe that the performance of KM and GMM is limited by the fact that the training set does not contain many examples of such roofs. In contrast, the proposed approach uses synthetic models and so is not affected by a small set of training examples.

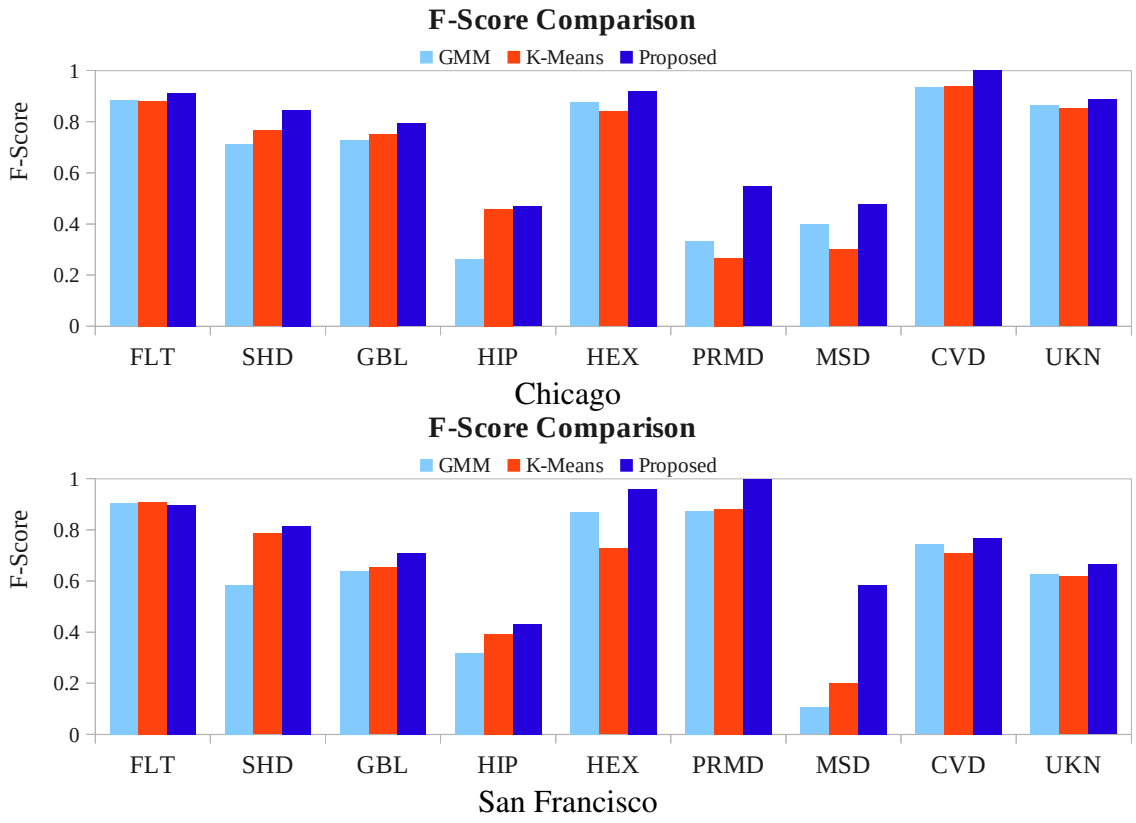


Figure 3.14. The comparison of F-Score on the two datasets by running KM, GMM and the proposed approach.

The confusion matrices of the proposed approach on the two datasets is shown in Figure **??**. As can be observed, the proposed approach performs well in classifying most of the roof styles. Errors mostly come from the confusion between the hip and gable roof styles, which are actually similar. Indeed, visual examination confirms that in my dataset many hip roofs resemble gable roofs due to shape erosion.

The hardest part in the classification is the recognition of the unknown roof styles, because there is no regular pattern (bag of words features) for roofs in this category. I observe that I get accuracy above 90% in the Chicago dataset and accuracy above 67% in the San Francisco dataset.

I evaluate the performance of the proposed approach as a function of the training set size. The F-Score obtained by the proposed approach as a function of training set size

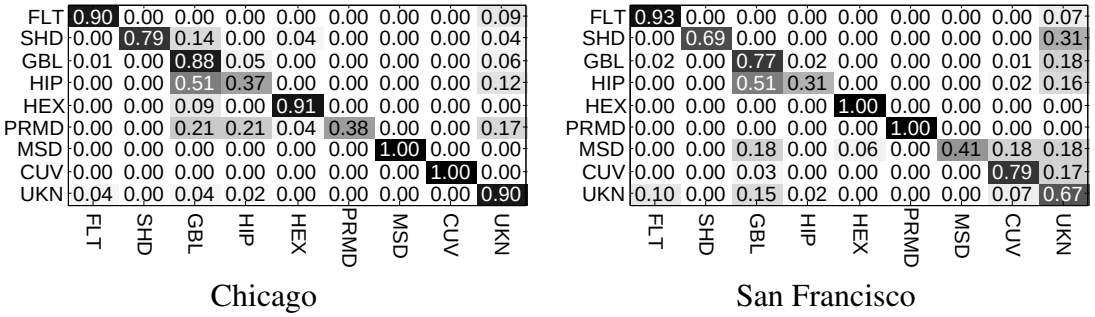


Figure 3.15. The confusion matrix of obtained by the proposed approach on two datasets.

is shown in Figure **??**.

As can be observed, the proposed approach achieves stable performance on most of the roof styles. For roof styles with relatively low score, the curves present are ascending which could suggest that a better score can be obtained once more training data is included.

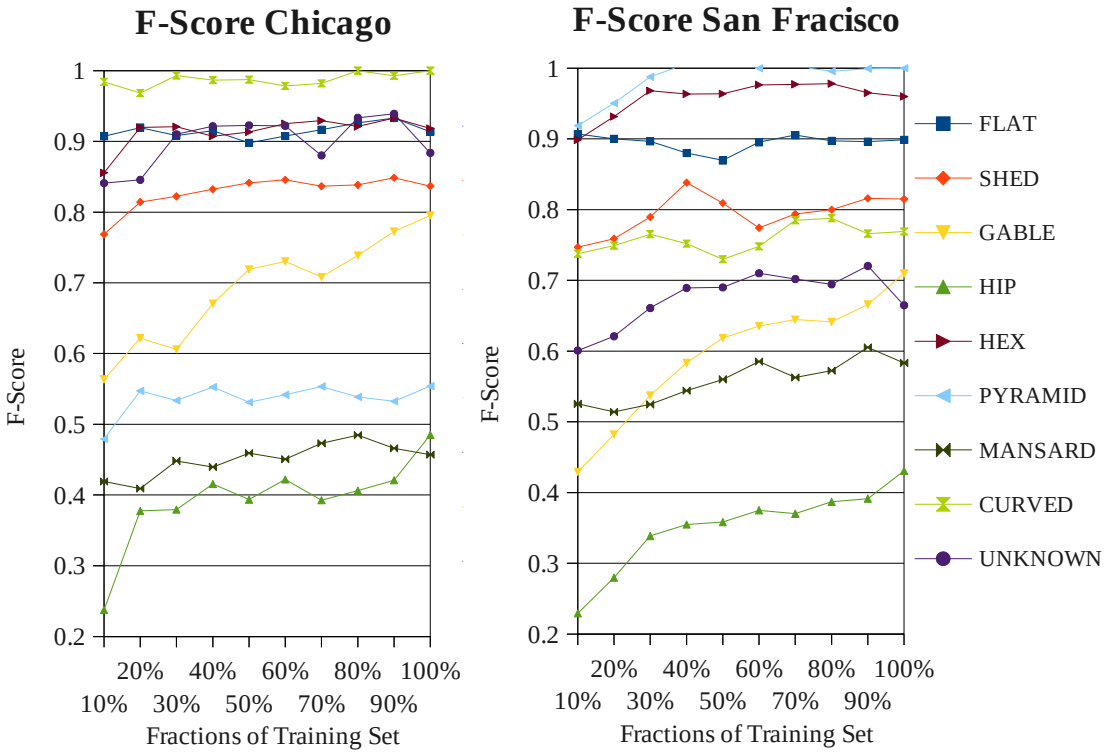


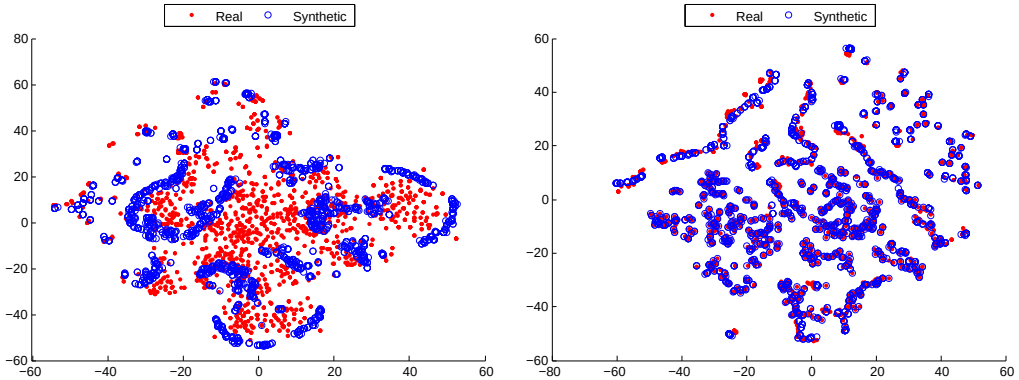
Figure 3.16. The F-Score of each roof style obtained by using an increasing proportion of the training data.

CHAPTER 4

ELIMINATION OF SYNTHETIC GAP

4.1 Introduction

Learning a classifier from synthetic data is unfortunately extremely challenging due to the following reasons. Firstly, the feature distribution of synthetic data generated will shift away from that of real data. Such distribution shift is termed synthetic gap and illustrated in Fig **??**. The synthetic gap is a major obstacle in using synthetic data to help learning classifiers, since synthetic data may fail to simulate the potential useful patterns of real data for training classifiers. To my knowledge, this synthetic gap problem has never been formally identified nor addressed in the literature. Secondly, since practically a small amount of labeled images may be available, it is necessary to jointly learn from synthetic and real data. The learning process must be automatically leveraged between synthetic data and real data.



(a) (b)

Figure 4.1. t-SNE visulization of synthetic gap using the data from SRC dataset. (a) synthetic gap of real and synthetic data; (b) MCAE bridges the synthetic gap.

To better learn a classifier from synthetic data, I propose a novel framework – Multichannel Autoencoder (MCAE) which is an extension of sparse autoencoder. The training step of MCAE is a process of bridging the synthetic gap between the real and the synthetic data by learning the mapping from (1) synthetic to real data and (2) real to real

data. Critically, such mapping try to keep the real data while enforce MCAE to learn a transfer from the synthetic data to the real data. I thus can generate more synthetic data which will simulate the real data when the learned mapping is applied to them.

4.2 Related Work

A large number of domain adaptation methods have been proposed over the recent years, and here I focus on the most related ones. Multiple methods perform unsupervised domain adaptation by matching the feature distributions in the source and the target domains. Some approaches perform this by reweighing or selecting samples from the source domain [**?**] [**?**] [**?**], while others seek an explicit feature space transformation that would map source distribution into the target ones [**?**] [**?**] [**?**]. An important aspect of the distribution matching approach is the way the (dis)similarity between distributions is measured. Here, one popular choice is matching the distribution means in the kernel-reproducing Hilbert space [**?**] [**?**], whereas [**?**] [**?**] map the principal axes associated with each of the distributions. my approach also attempts to match feature space distributions, however this is accomplished by modifying the feature representation itself rather than by reweighing or geometric transformation. Also, my method uses (implicitly) a rather different way to measure the disparity between distributions based on their separability by a deep discriminatively-trained classifier.

Several approaches perform gradual transition from the source to the target domain [**?**] [**?**] by a gradual change of the training distribution. Among these methods, [**?**] does this in a deep way by the layerwise training of a sequence of deep autoencoders, while gradually replacing source-domain samples with target-domain samples. This improves over a similar approach of (Glorot et al., 2011) that simply trains a single deep autoencoder for both domains. In both approaches, the actual classifier/predictor is learned in a separate step using the feature representation learned by autoencoder(s). In contrast to [**?**] [**?**], my approach performs feature learning, domain adaptation and classifier learning

jointly, in a unified architecture, and using a single learning algorithm (backpropagation). I therefore argue that my approach is simpler (both conceptually and in terms of its implementation). my method also achieves considerably better results on the popular OFFICE benchmark. While the above approaches perform unsupervised domain adaptation, there are approaches that perform supervised domain adaptation by exploiting labeled data from the target domain. In the context of deep feed-forward architectures, such data can be used to fine-tune the network trained on the source domain [**?**] [**?**] [**?**]. Finally, a recent and concurrent report by [**?**] also focuses on domain adaptation in feed-forward networks. Their set of techniques measures and minimizes the distance of the data means across domains. This approach may be regarded as a first-order approximation to my approach, which seeks a tighter alignment between distributions.

Autoencoder is one type of neural network and its output vectors have the same dimensionality as the input vectors [**?**]. The hidden representation obtained by training a sparse autoencoder followed by a parameters fine tuning is useful in pre-training a deeper neural network. Recently autoencoder with its different variants [**?**, **?**] also exhibit the success in learning and transferring sharing knowledge among data source from different domains [**?**, **?**, **?**], thus benefit other machine learning tasks.

Transfer Learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. Transfer learning has been found helpful in many real world problems, such as in sentiment classification [**?**], web page classification [**?**] and zero-shot classification of image and video data [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. Transfer learning is categorized to three classes [**?**]: inductive transfer learning, transductive transfer learning and unsupervised transfer learning. The work in this paper falls into a framework of domain adaptation [**?**, **?**] in the transductive transfer learning. Nonetheless, different from previous domain adaptation tasks of different source and target domains, the synthetic gap is caused by the shifted feature distribution of synthetic data from real data. To solve this problem,

my MCAE is developed from the idea of autoencoder.

4.3 Multichannel Autoencoder (MCAE)

In this section, I introduce the MCAE model as illustrated in Fig. **??**. It can (1) bridge synthetic gap by minimizing the discrepancy between real and synthetic data; and (2) preserve and emphasize the potential useful patterns existed in both real and synthetic data in order to generate the better feature representations used for learning classifiers.

Essentially, synthetic and real data should have similar patterns, a natural idea of bridging synthetic gap is to learning a mapping from the synthetic data to the real data using an autoencoder, and vice versa. MCAE, hence, provides a more flexible way to learn this mapping due to the specific structure of the MCAE. There are two channels in MCAE, left one and right one. Each channel basically is an SAE, however, two channels share the same hidden layer. With this structure, MCAE basically learns two tasks in the same time. By setting different types of input and out data such as the one in denoising autoencoder [**?**], MCAE is capable for many applications. In my work, to bridge the gap between synthetic data and real data, I set the task in left channel as one that takes synthetic data as input and real data as *reconstruction target*, while the task in right channel use real data in both input and *reconstrution target*. This configuration actually is essentially meaningful that by keeping the *reconstruction target* identical in two channels, MCAE attempts to transform inputs in two channels towards the same target, thus minimize the discrepancy between two input dataset which are synthetic data and real data in my work.

4.3.1 Problem setup. My MCAE is built on the sparse autoencoder (SAE). A basic autoencoder is a fully connected neural network with one hidden layer and can be decomposed into two parts: an encoding and a decoding process. Assume an input dataset with n instances $X = \{x_i\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^m$ and m is the dimension of each instance. Encoding typically transforms input data to hidden layer representation using an affine mapping

squashed by a sigmoid function:

$$h_e(x_i) = f(W_e x_i + b_e) \tag{4.1}$$

where $f(\cdot)$ is a sigmoid function and $\theta_e = \{W_e, b_e\}$, $W_e \in \mathbb{R}^{k \times m}, b_e \in \mathbb{R}^k$ is a set of unknown parameters in encoding with k nodes in hidden layer.

While in decoding, with parameters $\theta_d = \{W_d, b_d\}$, $W_d \in \mathbb{R}^{m \times k}, b_d \in \mathbb{R}^m$, autoencoder attempts to reconstruct the input data at the output layer by imposing another affine mapping followed by nonlinearity to hidden representation $h_e(x_i)$:

$$h_d(x_i) = f(W_d h_e(x_i) + b_d) \tag{4.2}$$

In above equation $h_d(x_i)$ is viewed as a reconstruction of input x_i. Normally, I impose $h_d(x_i) \approx x_i$. Here x_i play a role of *reconstruction target* in this expression and I use notation $\langle \mathfrak{i}{:}X_i, \mathfrak{t}{:}X_i \rangle$ to denote the configuration of input data short for $\mathfrak{i}$ and *reconstruction target* short for $\mathfrak{t}$ in an autoencoder. X_s and X_r indicate synthetic and real data respectively. By minimizing the reconstruction errors of all data instances, I have following objective function:

$$J(\theta_e, \theta_d) = \frac{1}{n} \sum_{i=1}^{n} (h_d(x_i) - x_i)^2 + \lambda W \tag{4.3}$$

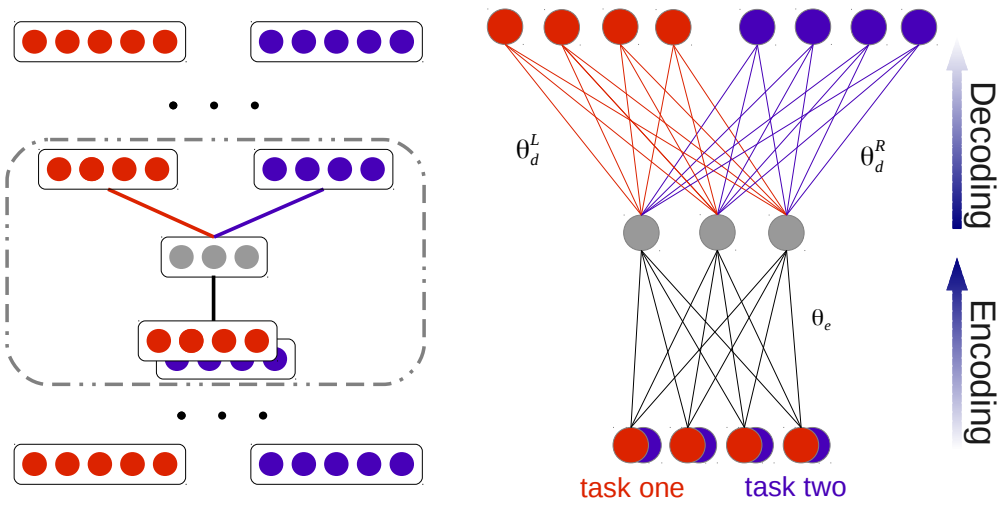
where $W = (\sum W_e^2 + \sum W_d^2)/2$ is a weight decay term added to improve generalization of the autoencoder and λ leverages the importance of this term.

To avoid learning identity mapping in autoencoder, a regularization term $\Theta = \sum_{i=1}^{k} \delta \log \frac{\delta}{\hat{\delta}_i} + (1-\delta) \log \frac{1-\delta}{1-\hat{\delta}_i}$ that penalizes over-activation of the nodes in the hidden layer is added. δ is a sparsity parameter and is set by users and $\hat{\delta}_i = \frac{1}{k} \sum_{i=1}^{k} h_e(x_i)$.

$$J(\theta_e, \theta_d) = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2 + \lambda W + \rho\Theta \tag{4.4}$$

ρ controls sparsity of representation in hidden layer.

Note that directly applying sparse autoencoder to my problem does not work well. For example, I can train an autoencoder purely by placing synthetic data in input layer and real data in output layer denoted as $\langle \text{i}{:}X_s,\ \text{t}{:}X_r \rangle$ which however can not bridge the synthetic gap in my problem. Such way of reconstruction is only to complement the missing information in synthetic data from real data but not vice versa



(a) (b)

Figure 4.2. (a) Illustration of the proposed MCAE model in a stacked autoencoder structure, where black edge between two layers are linked to and shared by two tasks, red and blue links are separately connected to left and right task respectively. (b) A zoom in structure of MCAE.

A better representation should be reconstructed by using the information from both real and synthetic data simultaneously. Specifically, I aim at two tasks: one is $\langle \text{i}{:}X_s, \text{t}{:}X_r \rangle^L$ which reconstructs synthetic data towards real data, and the other one is $\langle \text{i}{:}X_r,\ \text{t}{:}X_r \rangle^R$ which uses identical real data for input and *reconstruction target*, where $\langle\cdot\rangle^L$ and $\langle\cdot\rangle^R$ indicate the left and right channel of MCAE.

4.3.2 MCAE model. I propose a multichannel autoencoder that uses a balance regularization to leverage the learning between two tasks, i.e. $\langle \mathfrak{i}{:}X_s,\ \mathfrak{t}{:}X_r\rangle^L$ and $\langle \mathfrak{i}{:}X_r,\ \mathfrak{t}{:}X_r\rangle^R$. The structure of this new autoencoder is shown in Fig. **??**. In this new structure, tasks of two channels will share the same parameters θ_e in encoding process which will enforce autoencoder to reconstruct common structure in both tasks. However, in decoding process, I divide autoencoder to two separate channels that two tasks will have their own parameters θ_d^L and θ_d^R. Dividing autoencoder to two channels at decoding layer enable a more flexible control between the two tasks. Thus autoencoder better leverage the common knowledge from the two tasks.

With two channels in the MCAE, I target to minimize the reconstruction error of two tasks together while taking into account the balance between two channels. The new objective function of the MCAE is given in the following:

$$E = J^L(\theta_e, \theta_d^L) + J^R(\theta_e, \theta_d^R) + \gamma\Psi \tag{4.5}$$

where

$$\Psi = \frac{1}{2}(J^L(\theta_e, \theta_d^L) - J^R(\theta_e, \theta_d^R))^2 \tag{4.6}$$

is a regularization added to balance the learning rate between two channels. This regularization will have two effects on the MCAE. First, Ψ accelerates the speed of optimizing Eq. **??**, since minimizing Ψ requires both $J^L(\theta_e, \theta_d^L)$ and $J^R(\theta_e, \theta_d^R)$ are small which in turn cause E decreases faster. Second, Ψ penalizes a situation more when difference of learning error between two channels are large, so as to avoid imbalanced learning between two channels.

The minimization of Eq. **??** is achieved by back propagation and stochastic gradient descent using Quasi-Newton method. Since the regularization term is added to leverage the balance of different tasks, I have to compute the gradient of parameters θ_e and θ_d^L, θ_d^R in

MCAE.

With two branches in the MCAE, I target to minimize the reconstruction error of two tasks together while taking into account the balance between two branches. The new objective function of the YMAE is given in the following:

$$E = J^L(\theta_e, \theta_d^L) + J^R(\theta_e, \theta_d^R) + \gamma\Psi \tag{4.7}$$

where

$$\Psi = \frac{1}{2}(J^L(\theta_e, \theta_d^L) - J^R(\theta_e, \theta_d^R))^2 \tag{4.8}$$

is a regularization added to balance the learning rate between two branches. This regularization will have two effects on the YMAE. First, Ψ accelerates the speed of optimizing Eq. **??**, since minimizing Ψ requires both $J^L(\theta_e, \theta_d^L)$ and $J^R(\theta_e, \theta_d^R)$ are small which in turn cause E decrease faster. Second, Ψ penalize a situation more when difference of learning error between two branches are large, so as to avoid imbalanced learning between two branches.

The minimization of Eq. **??** is achieved by back propagation and stochastic gradient descent using Quasi-Newton method. In the MCAE, with balance regularization added to the objective, the only difference as opposed to sparse autoencoder is the gradient computation of unknown parameters θ_e and θ_d^L, θ_d^R. I clarify these differences in the following equations:

$$\begin{aligned}
\nabla_{W_e} E =& \frac{\partial J^L}{\partial W_e} + \frac{\partial J^R}{\partial W_e} + \gamma(J^L - J^R)(\frac{\partial J^L}{\partial W_e} - \frac{\partial J^R}{\partial W_e}) \\
\nabla_{b_e} E =& \frac{\partial J^L}{\partial b_e} + \frac{\partial J^R}{\partial b_e} + \gamma(J^L - J^R)(\frac{\partial J^L}{\partial b_e} - \frac{\partial J^R}{\partial b_e})
\end{aligned} \tag{4.9}$$

and

$$
\begin{aligned}
\nabla_{W_d^L} E &= \frac{\partial J^L}{\partial W_d^L} + \gamma(J^L - J^R)\frac{\partial J^L}{\partial W_d^L} \\
\nabla_{b_d^L} E &= \frac{\partial J^L}{\partial b_d^L} + \gamma(J^L - J^R)\frac{\partial J^L}{\partial b_d^L} \\
\nabla_{W_d^R} E &= \frac{\partial J^R}{\partial W_d^R} + \gamma(J^L - J^R)(-\frac{\partial J^R}{\partial W_d^R}) \\
\nabla_{b_d^R} E &= \frac{\partial J^R}{\partial b_d^R} + \gamma(J^L - J^R)(-\frac{\partial J^R}{\partial b_d^R})
\end{aligned}
\tag{4.10}
$$

The exact form of gradients of θ_e and θ_d^L, θ_d^R varies according to different sparsity regularization Θ used in the framework.

4.3.3 The advantages of MCAE over alternative Configurations. My MCAE enforces autoencoder to learn useful class patterns from the two tasks simultaneously. Thus it helps with capturing a common structure of synthetic and real images. Another alternative way is to concatenate the input and target of the two tasks $\langle \mathfrak{i}{:}X_s X_r,\ \mathfrak{t}{:}X_r X_r \rangle$ for autoencoder. I annotate the usage of this autoencoder as Concatenate-Input Autoencoder (CIAE), since this autoencoder learns concatenated tasks at the same time. Such configurations however may result in an unbalanced optimization for these two tasks: the optimization process of one task will take over the process of the other one. It results in a biased reconstructed hidden layer of the autoencoder and thus a limited classification performance.

4.4 Evaluation

The proposed MACE has been evaluated on two datasets I introduced in Chapter **??**. The first dataset is Satellite Roof Classification (SRC) and the second dataset is handwritten digit dataset from UCI machine learning repository.

Synthetic data are created to highlight the potential useful pattern existed in real images. I have two stages of generating synthetic data. In the first stage, for each real

data used to train MCAE, a synthetic version that best matching appearance of the real data is generated; thus pairs of corresponding real and synthetic data can be used to train the MCAE. In the second stage, more synthetic data could be derived using synthetic data generated in the first stage by both interpolation and extrapolation. To distinguish the set of synthetic data used in these two stages, I use abbreviation *Syn I* and *Syn II* to represent them respectively.

4.4.1 Experiment Settings. I fix the configuration of MCAE as $\langle \mathfrak{i}{:}X_s,\ \mathfrak{t}{:}X_r \rangle^L$ (left channel) and $\langle \mathfrak{i}{:}X_r,\ \mathfrak{t}{:}X_r \rangle^R$ (right channel). Specifically, the left channel is the reconstruction process from synthetic data to real data, while the right channel works in the same way as a standard SAE. my experimental results will show that the representations learned in such way greatly benefit the performance of classifiers I compared.

In the experiments two different classifiers of utilizing learned representations from MCAE (from Sec. 3) are compared. In the first scenario, MCAE encodes input data to a representation (feature) in the hidden layer and a SVM using RBF kernel is employed in this case to show the performance of the classification. In the second scenario, MCAE takes the input images and produces the reconstructed images at the output layer. Features, in this case, are images, therefore can be fed to Convolutional Neural Network (CNN) for classification. In my experiments I build a LeNet-5 [**?**] which is originally created for digit recognition. I show that using the same number of input data, the performance of the CNN prefers to the data produced by the MCAE.

I summarize all evaluations and comparisons using F-1 score, which is defined as:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.11}$$

4.4.2 Result. To better evaluate the performance of the proposed MCAE, I compare MCAE with Concatenate-Input Autoencoder (CIAE) [**?**] and Sparse Autoencoder (SAE)

[**?**]. In these experiments, I evaluate the performance on two classifiers: a CNN using reconstructed images and SVM using encoded hidden layer representation. I present the results of these comparisons in Table **??** and Table **??** for SRC and handwritten digit datasets respectively. It could be observed from these two tables that although the performance of the CIAE is close to MCAE, the proposed MCAE gets a better performance almost in all the comparisons.

Table 4.1. F1-score of roof style classification using reconstructed images (in CNN) and encoded image features (in SVM). Second column shows the data used to train the autoencoder in the first column. In classification, Real+*Syn II* are used in the training of CNN and SVM. ; *Syn I* + Real means that I use concatenation of the Syn I and real images as the input for the corresponding autoencoders.

	Data to train autoencoder	CNN Reconstructed	SVM Encoded
MCAE	$\langle$i:*Syn I*, t:Real$\rangle^L$ $\langle$i:Real, t:Real$\rangle^R$	**0.68**	**0.80**
CIAE	$\langle$i:*Syn I* + Real, t:*Syn I* + Real$\rangle$	0.68	0.78
SAE	$\langle$i:*Syn I*, t:*Syn I*$\rangle$	0.63	0.59
SAE	$\langle$i:Real, t:Real$\rangle$	0.62	0.62

Table 4.2. F1-score of handwritten digit recognition.

	Data to train autoencoder	CNN Reconstructed	SVM Encoded
MCAE	$\langle$i:*Syn I*, t:Real$\rangle^L$ $\langle$i:Real, t:Real$\rangle^R$	**0.98**	**0.96**
CIAE	$\langle$i:*Syn I* + Real, t:*Syn I* + Real$\rangle$	0.97	0.96
SAE	$\langle$i:*Syn I*, t:*Syn I*$\rangle$	0.94	0.91
SAE	$\langle$i:Real, t:Real$\rangle$	0.95	0.65

To qualitatively show actual images and synthetic images do look more similar after being processed by MCAE, examples of roof style images are shown in Figure **??**.

Synthetic data help learning a better classifier. I designed another group of experiments. In these experiments three different configurations of data are either reconstructed and encoded using the proposed MCAE, then used to train a CNN or a SVM in the experiments. All results from these experiments are compared in Table **??** and Table **??** respectively. An interesting thing to notice is that in experiments, using synthetic data can

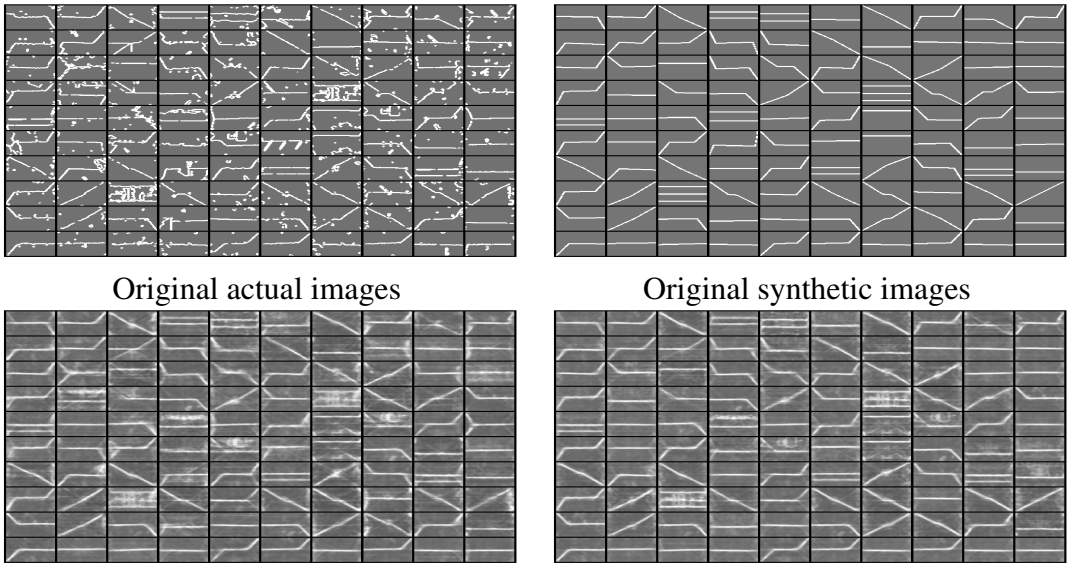


Figure 4.3. Examples of images before and after being reconstructed by MCAE. It could be observed from the images that actual images and synthetic images look much more similar after being processed by MCAE.

only achieve the same result as using a combination of real and synthetic data. This result proves that the distribution of the real data in this case is almost overlapping with the distribution of the synthetic data.

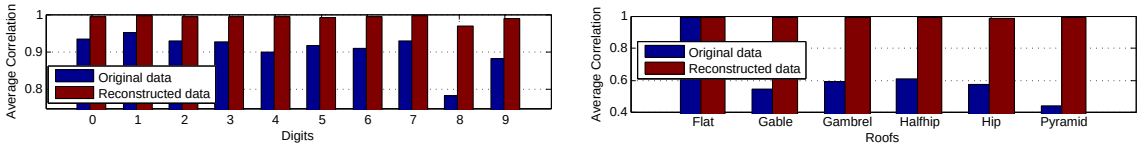


Figure 4.4. Correlation between real and corresponding best matching *Syn I* data.

MCAE bridges the synthetic gap. I compare the correlation defined as:

$$\text{Corr} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)\text{Var}(Y)} \tag{4.12}$$

between real and *Syn I* data before and after being reconstructed by the MCAE. The intention of these comparisons is to show that real synthetic images become much more alike each other in terms of the appearance after being reconstructed by the MCAE. The results

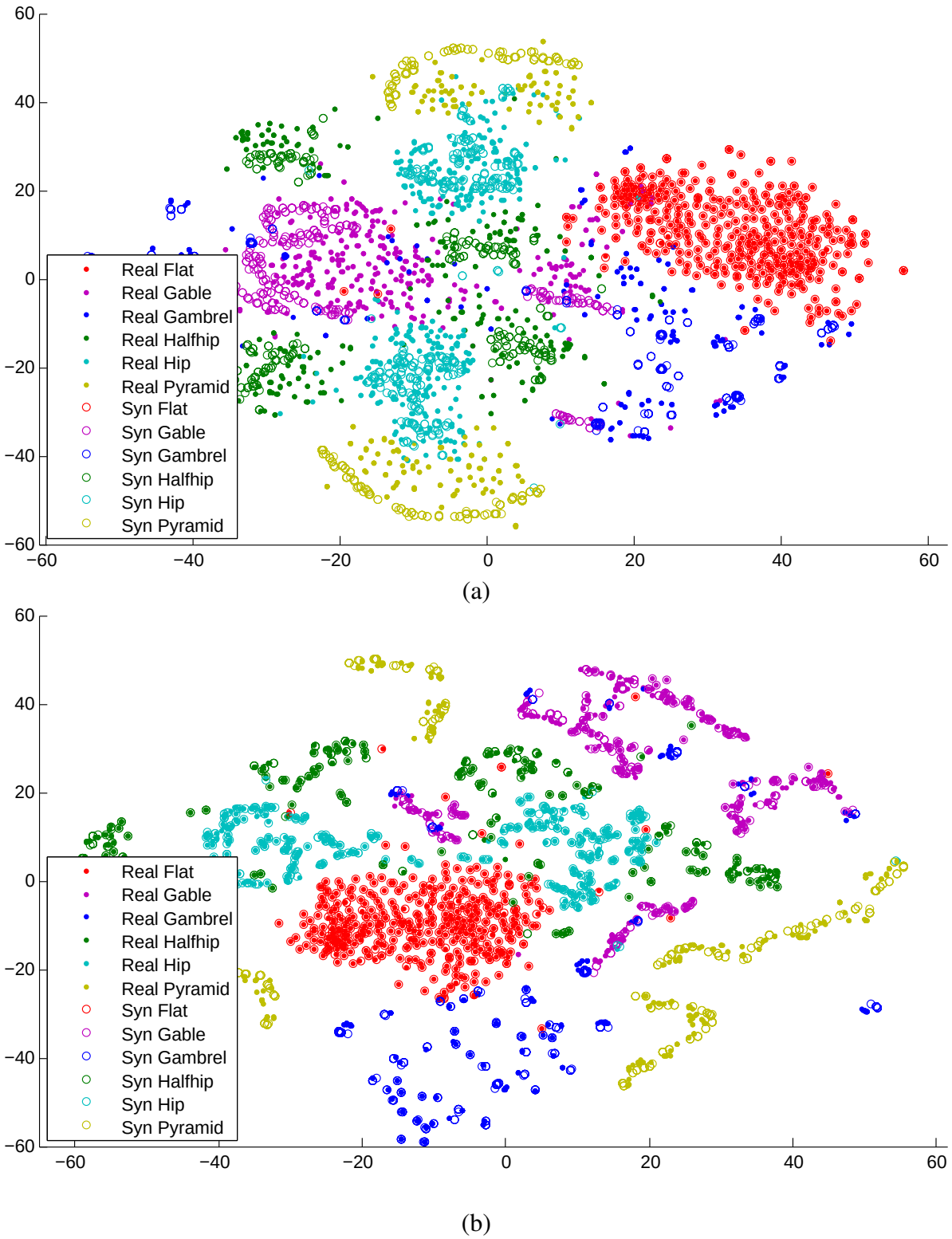


Figure 4.5. t-SNE [**?**] visualization of synthetic gap bridged by MCAE. (a) Data distributions of each class of SRC dataset. For many data instances, the (circle) real and (dot points) synthetic data are not overlapping. This is synthetic gap. (b) Data distributions of the reconstructed images by MCAE for each class of SRC dataset. The reconstructed images of all the real (circle) and synthetic (dot points) are almost overlapped. It means that my MCAE can bridge the synthetic gap.

Table 4.3. F1-score of roof style classification by classifier (CNN and SVM) using different set of data reconstructed of encoded using the proposed MCAE.

	Feature type	Real	*Syn II*	Real+*Syn II*
CNN	Reconstructed	0.65	0.68	0.68
SVM	Encoded	0.77	0.78	0.80

Table 4.4. F1-score of handwritten digit recognition.

	Feature type	Real	*Syn II*	Real+*Syn II*
CNN	Reconstructed	0.94	0.96	0.96
SVM	Encoded	0.96	0.96	0.98

are shown in Fig. **??**. It is shown that my method almost achieves 100% correlation between real and *Syn I* when both data are reconstructed by the proposed MCAE. That means the proposed MCAE bridges the synthetic gap between the real data and the synthetic data. The results are shown in Fig. **??**. It intuitively shows that my MCAE can help bridge the synthetic gap between real and synthetic data.

CHAPTER 5

CREATION OF SYNTHETIC DATA IN FEATURE SPACE

5.1 Introduction

In many real world problems, the distribution of data between classes is imbalanced. Learning from imbalanced datasets is an important research problem with many applications.

The fundamental issue in imbalanced learning is the ability of imbalanced data to significantly compromise the performance of standard learning algorithms [**?**]. Generally, there are three primary reasons that can cause this problem [**?**].

The first reason is that the lack of data in the minority class makes it difficult to detect regularities within the minority class. Thus, the learned decision boundaries are less likely to approximate the true decision boundaries.

Second, many classification algorithms utilize a general bias for better generalization and to avoid overfitting during learning. However, such bias can adversely affect the ability to learn the minority class. Inductive bias also plays a key role with respect to the minority class. Most classification algorithms prefer more common classes in the presence of uncertainty (i.e., they are biased in favor of the class priors).

Last but not least, noise exerts a greater impact on the minority class, because in this case it is more difficult for a classifier to distinguish noise from minority data. This is especially so in extreme cases where the number of noisy samples is greater than actual minority samples. The problem of overfitting rises again, when modifying the classifier to learn the minority data correctly.

To address these problems, numerous research efforts have been devoted to imbalanced learning in recent years. The majority of techniques that solve the imbalanced learn-

ing problem fall into two categories: cost-sensitive methods and sampling-based methods. In the next section, we review related work on sampling-based methods.

5.1.1 Related work. A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels [**?**]. There are mainly three groups of methods that can solve imbalanced learning problem [**?**] including sampling methods, cost sensitive methods, and kernel methods. Sampling-based methods are very effective and easy to use when solving imbalanced learning problems. In addition, sampling-based methods can be used together with methods in the other two groups to further improve performance. In such approaches a sampling technique is used to modify an imbalanced dataset to produce a balanced distribution. It has been shown that for most imbalanced datasets, sampling techniques do improve classification accuracy.

The basic sampling methods include undersampling and oversampling. Undersampling reduces majority class samples while oversampling increases minority class samples. While several works achieving data balance through undersampling have been proposed in the past [**?**][**?**], more research efforts have been devoted to oversampling due to the fact that oversampling does not discard information.

The simplest form of oversampling is duplication of minority class samples. This approach decreases the overall level of class imbalance, but may lead to overfitting [**?**]. SMOTE [**?**] is a fundamental approach for oversampling using data synthesis. To balance the dataset, SMOTE randomly selects a seed sample and synthesizes a new sample by applying a linear interpolation between the seed sample and one of its neighbors. Large research efforts have been devoted to feature space data synthesis based on SMOTE. Several methods integrate data synthesis as a part of the learning procedure. For example, by introducing SMOTE in each iteration of boosting, SMOTEBoost [**?**] increases the number of minority class samples and focus on these cases in each boosting iteration. Using the same idea of boosting, DataBoost-IM [**?**] and RAMOBoost [**?**] discover samples difficult

to classify during each iteration of boosting, which are used to guide the oversampling in both the majority and minority classes.

In another group of minority oversampling approaches, the data synthesis procedure is independent of the learning processes. Such methods give preferences to different regions of a dataset by assigning weights to samples in the dataset. These weights can then generate a probability distribution which is used for randomly drawing samples. In such approaches the data synthesis can be completed in one step. Methods in this group include Borderline-SMOTE [**?**], Adasyn [**?**], [**?**] and MWMOTE [**?**]. All of these methods synthesize more samples along decision boundaries. However, these methods do not have objective functions to systematically guide the process of oversampling and so do not have a systematic way to decide on where new data should be synthesized. Thus, such approaches cannot measure the impact of each synthetic sample. As a result, there are several potential problems. One is that the oversampling procedure may sacrifice the performance of the majority class in order to improve the performance of the minority class in the classification. Another is that synthetic minority samples themselves can be misclassified and affect the performance in the minority class.

5.1.2 Novel Contribution. The proposed approach, CGMOS, is a member of the SMOTE family that can achieve data oversampling in a single step. To address some of the shortcomings in existing approaches, we propose a novel oversampling strategy by systematically considering the performance of both minority and majority classes. Based on a Bayesian classification framework, our proposed approach computes the influence of minority data addition on the certainty of the entire dataset. CGMOS thus can synthesize new samples that will improve the overall certainty of the entire dataset in classification. We prove that during training CGMOS is guaranteed to perform better than SMOTE when using Bayesian classification. To validate the proof, We further show experimentally that CGMOS outperforms known approaches when tested on real-world data set collections

using different classifiers.

5.2 Problem Formulation

In this paper, we address the binary classification problem for imbalanced datasets. Let $D = \{(x_j, y_j)\}_{j=1}^{n}$ be a training dataset, where $x_j \in \mathfrak{R}^m$ are features and $y_j \in \{l_{\text{mjr}}, l_{\text{mnr}}\}$ are ground truth class labels. We begin by formally defining the certainty of imbalanced binary classification using a Bayesian framework, where a kernel density estimation (KDE) is used to estimate the samples' probability density function (PDF). We then show how CGMOS can synthesize more samples according to the certainty estimation.

5.2.1 Definition of Certainty. Suppose (x_j, y_j) is any tuple in the training dataset D, where x_j is a feature vector and y_j is the ground truth label of x_j.

A Bayesian classifier maps $x_j \rightarrow l, l \in \{l_{\text{mjr}}, l_{\text{mnr}}\}$ using following rule.

$$l = \begin{cases} l_{\text{mnr}} & \text{if } \frac{P(l_{\text{mnr}}|x_j)}{P(l_{\text{mjr}}|x_j)} > 1 \\ l_{\text{mjr}} & \text{otherwise} \end{cases}$$

where the posterior probability $P(l \mid x_j)$ is computed using Bayes' rule:

$$P(l \mid x_j) = \frac{P(x_j \mid l)P(l)}{P(x_j)}; \qquad l \in \{l_{\text{mjr}}, l_{\text{mnr}}\}$$

Uncertainty is commonly used in machine learning algorithms. In this work, we use the posterior probability $P(y_j \mid x_j)$ to define certainty. This is because in classification, the posterior probabilities $P(y_j \mid x_j)$ reflect the certainty of assigning a sample to a correct label, where higher numbers indicate classification results with a stronger certainty.

Definition 1. (Certainty) Let (x_j, y_j) be any tuple in D, where x_j is a feature vector and

y_j is the ground truth label of x_j. The certainties for samples in the majority and minority class are respectively defined as:

$$C(y_j = l_{\text{mjr}} \mid x_j) = P(y_j = l_{\text{mjr}} \mid x_j) \tag{5.1}$$

$$C(y_j = l_{\text{mnr}} \mid x_j) = P(y_j = l_{\text{mnr}} \mid x_j) \tag{5.2}$$

It should be noted that in the case of binary classification the definition of certainty above is related up to some constants to the uncertainty defined in [**?**] based on margin confidence.

5.2.2 PDF Estimation. There are two general ways to estimate a density function: parametric or non-parametric. In this work we use a non-parametric model so as to not depend on a specific distribution model. We use kernel density estimation (KDE)[**?**][**?**] to estimate the likelihood $P(x_j \mid l)$, $l \in \{l_{\text{mjr}}, l_{\text{mnr}}\}$.

Assuming that the data is independent and identically distributed (i.i.d) and drawn from some distribution with an unknown density $P(x_j \mid l)$, we have using KDE:

$$P(x_j \mid l) = \frac{\sum_{k=1}^{n} K(\frac{x_j - x_k}{h_k}) \mathrm{I}(y_k = l)}{\sum_{k=1}^{n} \mathrm{I}(y_k = l)} \tag{5.3}$$

where $l \in \{l_{\text{mjr}}, l_{\text{mnr}}\}$, $I(\cdot)$ is an indicator function, and $K(\cdot)$ is a kernel function which has zero mean and integrates to one. Given any sample x_k, the bandwidth h_k of the sample x_k controls the effective range of the kernel and smoothness of the density function. Intuitively one wants to choose h_k as small as the data allows to exhibit as many underlying structures of the data as possible. Small bandwidth, however, will result in a noisy estimate. In this work, for any sample x_k, we calculate a bandwidth h_k as a scaled average distance

between x_k and its q nearest neighbors:

$$h_k = \sigma \cdot \frac{\sum_{x \in N(x_k)} \|x - x_k\|}{q} \tag{5.4}$$

where $N(x)$ is the set of the q nearest neighbors of x_k and $\sigma > 0$ is a scale factor applied to the distance.

5.2.3 Oversampling Seed Selection. In most classification algorithms, samples close to decision boundaries have less certain classification results. In order to achieve better predictions for such samples, many existing approaches synthesize data directly along the boundaries. However, this is risky and the expected performance improvement is not guaranteed. There are two primary reasons. First, samples from both classes are mixed in regions near the boundaries. Synthetic samples if added to these regions are less predictable and hard to learn. Second, adding synthetic minority samples to these regions may adversely impact the majority class, which may in turn decrease the performance of the majority class in classification. Instead of unguided oversampling near the boundaries, our proposed approach targets adding samples by considering the certainties of both the minority and majority classes before and after adding the samples. The synthetic samples thus are added to locations that can improve the overall certainty of the original data and boost the performance of the classification.

CGMOS uses a similar procedure as SMOTE when synthesizing a new sample. The sample is produced by interpolating between one seed sample and some of its neighbors. However, instead of randomly drawing a seed sample for interpolation, CGMOS assigns each sample $(x_i, y_i) \in D$ a weight $W(x_i)$ which is used to determine the probabilities of x_i being chosen for interpolation. A higher weight results in a higher probability of a point being selected.

To compute $W(x_i)$, we suppose that a new sample will be added to the same loca-

tion as x_i. The weight $W(x_i)$ is computed as a relative certainty change[1] comparing the certainty before and after the sample is added. With a new sample added at location x_i, we update the certainty for all $(x_j, y_j) \in D$ and denote it as $C_{+i}(y_j \mid x_j)$.

Definition 2. (Relative Certainty Change) The relative certainty change of label y_j assigned to feature x_j due to adding a minority example at location x_i is defined by:

$$R_{+i}(y_j \mid x_j) = \frac{C_{+i}(y_j \mid x_j) - C(y_j \mid x_j)}{C(y_j \mid x_j)} \tag{5.5}$$

where $C(y_j \mid x_j)$ is the certainty before addition.

When computing $W(x_i)$, CGMOS considers the relative certainty changes of examples from both the majority and the minority classes. $W(x_i)$ is computed as the average value of relative certainty changes of all samples in the dataset.

$$W(x_i) = 1 + \frac{1}{n}\sum_{j=1}^{n} R_{+i}(y_j \mid x_j) \tag{5.6}$$

Given $W(x_i)$ for all $x_i \in D$, it is easy to see $W(x_i) > 0$. We compute a normalization factor z so that $\frac{1}{z}\sum_{i=1}^{n} W(x_i) = 1$. Therefore, the oversampling procedure can randomly choose sample for interpolation according $W(x_i)/z$. The interpolation phase of CGMOS is the same as SMOTE[**?**].

A demonstration of CGMOS is shown in Fig. **??**. In this figure, samples in both the majority and minority classes are randomly drawn based on Gaussian distribution, where the means of the two datasets are on the same horizontal line, and the mean of the majority

[1]Measuring absolute certainty increments will not work, because measuring magnitude will give higher preference to parts which already have high certainty.

is to the right of the minority. The majority class contains 2000 samples and the minority class contains 400 samples. Color in part 1 of the figure indicates the certainty of each example with respect to its class, where red indicates high certainty. We highlight 3 regions (A, B, C) in the minority class. Samples in region A have relative high certainties, sample in region B has low certainties and region C is a boundary region in which samples have the lowest certainties. Part 2 of the figure shows the weight of each example as computed by our approach where red indicates high values. Region B has higher values and is where CGMOS will synthesize most of the samples.

To show the certainty changes induced by adding samples at different locations of the dataset, in part 3 of the figure we add one minority sample and move its location with a fixed step size from left to right on a horizontal line passing through the two classes. We then compute the relative certainty changes for all samples in both classes. As can be observed, by measuring relative certainty changes, CGMOS will assign a higher weight to samples in region B. The figure also shows that by oversampling more in region B, the certainty of the entire dataset gets improved, because the relative certainty changes are positive.

5.3 Theoretical Guarantee Over SMOTE

Several existing approaches claim handling imbalanced learning better than SMOTE. Such claims are normally validated using empirical tests without a theoretical guarantee and in some instances may not extend to new datasets. In this section we provide a theoretical guarantee showing that CGMOS is expected to work better than SMOTE in training process.

Let $D = \{(x_j, y_j)\}_{j=1}^{n}$ be a training dataset. Let $W(D) = \{W(x_i)\}_{i=1}^{n}$ be the sample weights computed using Eqn. **??**.

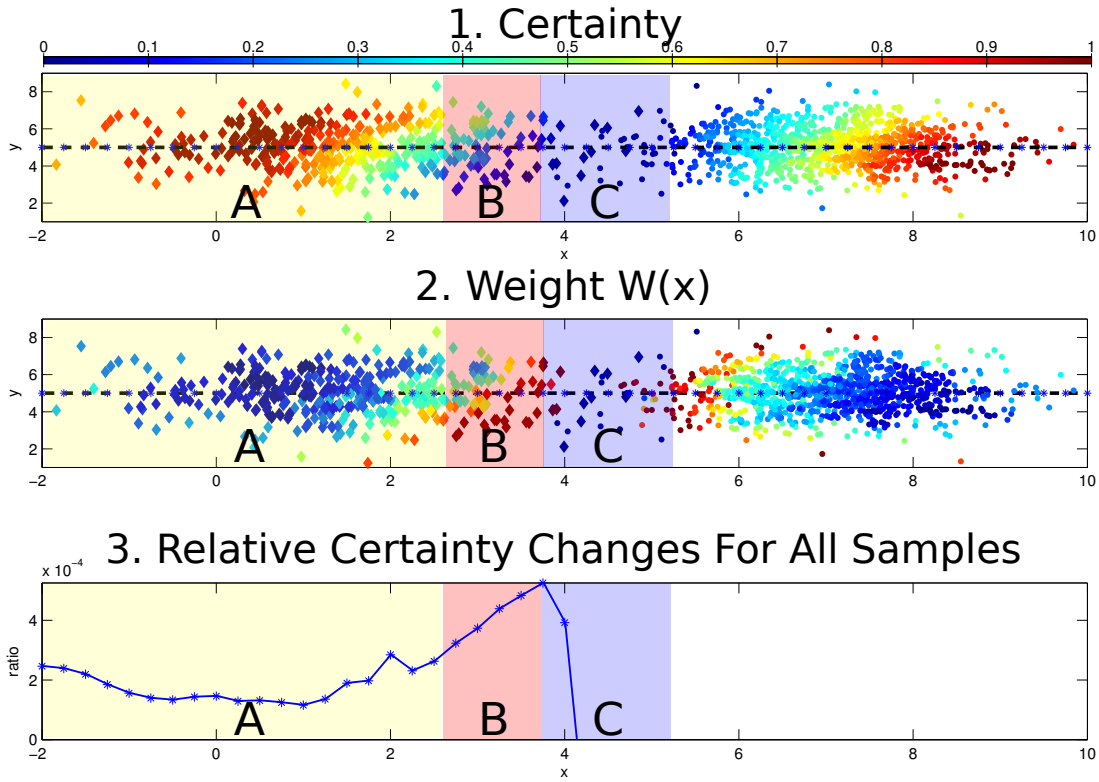


Figure 5.1. Demonstration of CGMOS. In first two figures, diamonds represent minority samples and circles represent majority samples. The positions of synthesized data points are labeled using a star symbol on a horizontal line passing through the center. The x and y axes represent features. In the bottom figure the x axis indicates a location where a sample was added (in correspondence with the first two figures) whereas the y-axis indicates the relative certainty change.

Lemma 1. *Given a set of weights* $\{W(x_i)\}_{i=1}^{n}$ *as defined above and a normalization factor* z *given by* $z = \sum_{i=1}^{n} W(x_i)$*, it must be that* $\sum_{i=1}^{n} W(x_i)^2 \geq \frac{z^2}{n}$.

Proof Let W be an n-dimensional vector whose elements are $W(x_i)$. Let I be an n-dimensional vector whose elements are all 1. Using the Cauchy-Schwarz inequality we have: $|\ W \cdot I| \leq \|W\| \cdot \|I\|$.

Definition 3. (Addition Likelihood Ratio) Let θ denote the non-parametric likelihood estimate $P(x_j \mid l)$, $l \in \{l_{\text{mjr}}, l_{\text{mnr}}\}$ before a new sample x_i is added, and θ' denote the non-parametric likelihood estimate after the new sample is added. The addition likelihood

ratio $r_{+i}(y_j \mid x_j)$ of example x_j by adding data to x_i location is defined as the ratio between the likelihood estimate after the new addition and the likelihood estimate before the new addition:

$$r_{+i}(y_j \mid x_j) \equiv P(y_j \mid x_j;\ \theta')/P(y_j \mid x_j;\ \theta). \tag{5.7}$$

Lemma 2. *The addition likelihood ratio $r_{+i}(y_j \mid x_j)$ is related to the relative certainty change ratio $R_{+i}(y_j \mid x_j)$ by:*

$$r_{+i}(y_j \mid x_j) = 1 + R_{+i}(y_j \mid x_j). \tag{5.8}$$

Proof According to the definition of the certainty, we have $C_{+i}(y_j \mid x_j) = P(y_j \mid x_j;\ \theta')$ and $C(y_j \mid x_j;\ \theta) = P(y_j \mid x_j;\ \theta)$. Then $P(y_j \mid x_j;\ \theta') = r_{+i}(y_j \mid x_j)P(y_j \mid x_j;\ \theta)$ according to the definition of likelihood ratio. Given Eqn. **??**, we have that $R_{+i}(y_j \mid x_j) = \frac{r_{+i}(y_j|x_j)P(y_j|x_j;\theta)-P(y_j|x_j;\theta)}{P(y_j|x_j;\theta)}$. By simplifying this equation, we thus have

$$r_{+i}(y_j \mid x_j) = 1 + R_{+i}(y_j \mid x_j). \quad \blacksquare \tag{5.9}$$

The addition likelihood ratio defined in Eqn. **??** measures the gain in adding a new point, where higher gains are desired. Note that while the gain is normally close to 1 it may be bigger or smaller than 1.

Definition 4. (Average gain) The average gain when adding sample x_i is defined by:

$$\bar{r}_{+i} = \frac{1}{n}\sum_{j=1}^{n} r_{+i}(y_j \mid x_j) \tag{5.10}$$

Lemma 3. *Given the average gain, it must be that:*

$$\bar{r}_{+i} = W(x_i). \tag{5.11}$$

Proof Using the definition of $W(x_i)$ we have $W(x_i) = \frac{1}{n}\sum_{j=1}^{n} R_{+i}(y_j \mid x_j)$. Using Lemma 2 we can replace $r_{+i}(y_j \mid x_j) - 1$ with $R_{+i}(y_j \mid x_j)$. Hence:

$$\bar{r}_{+i} = \frac{1}{n}\sum_{j=1}^{n} R_{+i}(y_j \mid x_j) + 1 \equiv W(x_i) \quad \blacksquare \tag{5.12}$$

The average gain is an indicator of the benefit of CGMOS. We show that the expected average gain is higher in proposed approach compared with SMOTE.

Theorem 1. *The expected average gain in CGMOS is higher or equal to that of SMOTE.*

Proof For CGMOS the expected average gain is given by:

$$E_p \equiv E[\bar{r}_{+i}] = \sum_{i=1}^{n} \bar{r}_{+i}\frac{W(x_i)}{z} \tag{5.13}$$

where z is the normalization factor as defined earlier. Using Lemma 3:

$$E_p = \sum_{i=1}^{n} W(x_i)\frac{W(x_i)}{z} = \frac{1}{z}\sum_{i=1}^{n} W^2(x_i). \tag{5.14}$$

For SMOTE the expected average gain is given by:

$$E_s \equiv E[\bar{r}_{+i}] = \sum_{i=1}^{n} \bar{r}_{+i}\frac{1}{n} \tag{5.15}$$

Using Lemma 3:

$$E_s = \frac{1}{n}\sum_{i=1}^{n} W(x_i) = \frac{z}{n} \tag{5.16}$$

Using Lemma 1:

$$E_p = \frac{1}{z}\sum_{i=1}^{n} W^2(x_i) \geq \frac{1}{z}\frac{z^2}{n} = E_s \quad \blacksquare \tag{5.17}$$

5.3.1 Datasets. 30 real-world datasets were randomly chosen from the UCI machine

learning repository [**?**] for empirical testing of CGMOS. Most of the datasets were released within the past 10 years. As some of the datasets contain samples of more than two classes, we convert such datasets to a binary classification problem by keeping the class with the least data and merging all other classes. A summary of the test collections is provided in Table **??**.

Table 5.1. Summary of the datasets used in our experiments, where S#, F#, and R stand for the number of samples, the number of features, and imbalance ratio (defined as #minority/#majority).

Name	S#	F#	R	Year	Name	S#	F#	R	Year
BankMarket	45211	17	0.13	2012	Libras	360	91	0.07	2009
BloodService	748	5	0.31	2008	MultipleFs	2000	649	0.11	1998
BreastCancer	400	9	0.53	1988	Parkinson	1040	26	0.02	2014
BreastTissue	106	10	0.15	2010	PlanRelax	182	13	0.4	2012
CarEvaluation	1730	6	0.04	1997	QSAR	1055	41	0.51	2013
Card'graphy	2126	23	0.09	2010	SPECT	268	22	0.26	2001
CharacterTraj	2860	3	0.04	2008	SPECTF	134	44	0.26	2001
Chess	3198	22	0.91	1989	SeismicBumps	2584	19	0.07	2013
ClimateSim	540	18	0.09	2013	Statlog	2310	19	0.17	1990
Contraceptive	1474	9	0.29	1997	PlatesFaults	1941	27	0.03	2010
Fertility	100	10	0.14	2013	TAEvaluation	151	5	0.49	1997
Haberman	306	3	0.36	1999	UKnowledge	403	5	0.1	2013
ILPD	580	10	0.4	2012	Vertebral	310	6	0.48	2011
ImgSeg	2310	19	0.17	1990	Customers	440	8	0.48	2014
Leaf	342	16	0.24	2014	Yeast	1484	8	0.04	1996

5.3.2 Compared Approaches. According to a survey of imbalanced learning [**?**], there are mainly three groups of methods addressing imbalanced learning: sampling methods, cost sensitive methods, and kernel methods. The proposed CGMOS belongs to the sampling group. Thus, we compare CGMOS to five other oversampling methods in this group: SMOTE[**?**], Borderline-SMOTE[**?**], ADASYN[**?**], MWMOTE[**?**] and RAMOBoost[**?**]. Since oversampling by duplication is broadly used in many applications as a baseline, we add it to our evaluation as well. To demonstrate the improvement of these oversampling strategies, we include in the comparison raw data with no oversampling. It should be noted that sampling methods are often combined with cost sensitive methods and kernel methods to

further boost learning. [**?**][**?**][**?**].

5.3.3 Base classifiers. We match the compared classifiers to classifiers used in other SMOTE extension evaluations. Six well-known classifiers are tested in experiments. The first is the Bayesian classifier based on kernel density estimation described in Section **??** (b-kde). The second is a K nearest neighbors classifier (knn). The third is a support vector machine classifier using RBF kernel (svm). The fourth one is a neural network (nn) with one hidden layer. We use in addition two ensemble methods: a random forest implementing the C4.5 decision tree [**?**] (rf) and Adaboost.M1 [**?**]. All hyper-parameters of the classifiers tested were determined by cross validation to ensure the best performance of each method.

5.3.4 Evaluation metric. Finding an appropriate evaluation metric for different tasks is challenging, since different evaluation metrics are designed for different purposes. The datasets used in this paper cover from financial application to medical treatment. To achieve an general evaluation and avoid bias, we follow the method in [**?**][**?**][**?**][**?**][**?**] and use different metrics to evaluate the performance of the proposed CGMOS oversampling algorithm.

Among these evaluation metrics, the most frequently adopted ones are $Precision$ and $Recall$ when the focus of evaluation is focus on one specific class such as problems in text classification, information extraction, natural language processing and bioinformatics. In these areas of application the number of examples belonging to one class is often substantially lower than the overall number of examples, which basically are imbalance learning problems. $Precision$ and $Recall$ are defined as:

$$Precision = \frac{TP}{(TP + FP)}$$
$$Recall = \frac{TP}{(TP + FN)}$$

However, these two metrics share an inverse relationship between each other. A quick inspection on the $Precision$ and $Recall$ formulas readily yields that solely use each of these two metrics only provide a limit view of an algorithm under test. As $Recall$ provides no insight to how many examples are incorrectly labeled as positive and $Precision$ cannot assert how many positive examples are labeled incorrectly. Specifically, the $F\text{-}score$ combines $Precision$ and $Recall$ as measure of the effectiveness of classification in terms of a ration of the weighted importance on either $Recall$ or $Precision$, which is defined as:

$$F\text{-}score = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}.$$

We use $\beta = 1$ to treat $Precision$ and $Recall$ equally in all evaluations. As a result, $F\text{-}score$ provides more insight into the functionality of a classifier.

As $F\text{-}score$ measures the harmonic mean of $Precision$ and $Recall$, we also compute $Gscore$ which is the geometric mean of $Precision$ and $Recall$ and is able to evaluate the degree of inductive bias in terms of a ratio of positive accuracy and negative accuracy [**?**].

$$G\text{-}score = \sqrt{Precision \cdot Recall}$$

As both $F\text{-}score$ and $G\text{-}score$ concentrate their measures on one class (positive examples) [**?**], to have a general way of comparing our test results, we altered the positive examples between the majority and minority classes when computing $F\text{-}score$ and $G\text{-}score$. Thus we show $F\text{-}score$ and $G\text{-}score$ for the majority and the minority classes separately.

Although, both $F\text{-}score$ and $G\text{-}score$ are great evaluation metrics, they are still less

Table 5.2. A summary of AUC, $Precision$, $Recall$, $F\text{-}score$ and $G\text{-}score$ of all competitors for the majority and minority classes produced by 6 classifiers on the artificial datasets.

		Minority				Majority			
	AUC	**Precision**	**Recall**	**Fscore**	**Gscore**	**Precision**	**Recall**	**Fscore**	**Gscore**
b-kde									
Original	0.797	0.139	0.033	0.054	0.068	0.830	**0.995**	**0.905**	**0.909**
Dup	0.733	0.385	0.454	0.417	0.418	0.869	0.742	0.801	0.803
SMOTE	0.807	0.488	0.705	0.577	**0.587**	0.833	0.644	0.726	0.733
B-SMOTE	0.774	0.258	0.456	0.330	0.343	0.846	0.671	0.748	0.754
MWMOTE	0.794	0.396	**0.754**	0.520	0.547	0.836	0.557	0.669	0.682
ADASYN	0.802	0.395	0.632	0.487	0.500	0.817	0.598	0.691	0.699
RAMOboost	0.748	0.358	0.343	0.350	0.350	0.860	0.822	0.841	0.841
CGMOS	**0.842**	**0.536**	0.517	**0.526**	0.526	**0.908**	0.815	0.859	0.860
knn									
Original	0.821	**0.701**	0.521	0.598	0.604	0.902	**0.942**	**0.922**	**0.9217**
Dup	0.810	0.519	0.732	0.607	0.616	0.921	0.818	0.867	0.868
SMOTE	0.827	0.506	**0.804**	0.621	0.638	0.925	0.805	0.861	0.863
B-SMOTE	0.811	0.494	0.736	0.591	0.603	0.927	0.790	0.853	0.856
MWMOTE	0.832	0.504	0.792	0.616	0.632	**0.928**	0.794	0.856	0.858
ADASYN	0.825	0.495	0.786	0.607	0.623	**0.929**	0.786	0.851	0.854
RAMOboost	0.827	0.540	0.684	0.604	0.608	0.918	0.847	0.881	0.881
CGMOS	**0.840**	0.544	0.766	**0.636**	**0.646**	0.925	0.842	0.882	0.883
svm									
Original	0.792	**0.632**	0.587	0.609	0.609	0.882	0.935	0.908	0.908
Dup	0.815	0.543	0.436	0.484	0.487	**0.981**	0.861	0.917	0.919
SMOTE	0.844	0.579	0.726	0.644	0.648	0.879	0.844	0.861	0.861
B-SMOTE	0.832	0.475	0.729	0.575	0.588	0.893	**0.959**	**0.924**	**0.925**
MWMOTE	0.830	0.547	0.647	0.593	0.595	0.880	0.884	0.882	0.882
ADASYN	0.827	0.536	0.654	0.589	0.592	0.880	0.755	0.813	0.815
RAMOboost	0.842	0.556	0.673	0.609	0.611	0.968	0.852	0.906	0.908
CGMOS	**0.864**	0.555	**0.788**	**0.651**	**0.661**	0.943	0.830	0.883	0.885
nn									
Original	0.801	**0.632**	0.412	0.499	0.510	0.892	**0.962**	**0.925**	**0.9258**
Dup	0.843	0.543	0.777	0.639	0.650	0.926	0.819	0.869	0.871
SMOTE	0.840	0.555	0.750	0.638	0.645	0.921	0.820	0.868	0.869
B-SMOTE	0.841	0.475	0.779	0.590	0.608	0.924	0.802	0.859	0.861
MWMOTE	0.841	0.547	0.778	0.642	0.652	0.927	0.812	0.866	0.867
ADASYN	0.842	0.536	**0.786**	0.637	0.649	0.929	0.803	0.861	0.863
RAMOboost	0.841	0.556	0.743	0.636	0.643	0.919	0.830	0.872	0.873
CGMOS	**0.865**	0.579	0.750	**0.653**	**0.659**	**0.933**	0.845	0.887	0.888
rf									
Original	0.872	**0.699**	0.534	0.606	0.611	0.909	**0.956**	**0.932**	**0.932**
Dup	0.873	0.682	0.641	0.661	0.661	0.917	0.924	0.921	0.921
SMOTE	0.875	0.667	0.655	0.661	0.661	0.920	0.917	0.918	0.918
B-SMOTE	0.867	0.653	0.637	0.645	0.645	0.920	0.906	0.913	0.913
MWMOTE	0.878	0.658	0.651	0.655	0.655	0.920	0.922	0.921	0.921
ADASYN	0.876	0.663	0.669	0.666	0.666	0.919	0.915	0.917	0.917
RAMOboost	0.874	0.686	0.618	0.650	0.651	0.915	0.933	0.924	0.924
CGMOS	**0.884**	0.685	**0.678**	**0.681**	**0.681**	**0.923**	0.926	0.924	0.924
Adaboost.M1									
Original	0.868	**0.699**	0.572	0.629	0.632	0.906	**0.944**	**0.925**	**0.9247**
Dup	0.865	0.622	0.708	0.662	0.664	0.922	0.873	0.897	0.897
SMOTE	0.867	0.608	0.714	0.657	0.659	0.923	0.880	0.901	0.901
B-SMOTE	0.864	0.581	0.724	0.644	0.648	**0.927**	0.861	0.893	0.893
MWMOTE	0.868	0.600	0.708	0.650	0.652	0.922	0.880	0.901	0.901
ADASYN	0.867	0.599	0.726	0.657	0.660	0.925	0.873	0.898	0.899
RAMOboost	0.865	0.631	0.699	0.663	0.664	0.922	0.882	0.901	0.902
CGMOS	**0.871**	0.619	**0.728**	**0.670**	**0.672**	0.925	0.882	0.903	0.903

effective in some situations. So we also employ the ROC graphs [**?**][**?**][**?**] in the evaluation. ROC graph is a two-dimensional graph, while $FP\ rate$ and $TP\ rate$ are its X axis and Y axis respectively.

An ROC graph basically manifest its usefulness by showing relative trade-off between benefit (true positive) and cost (false positive). One attractive property make ROC graph a good metric in imbalanced learning lies in the facts that ROC curve is insensitive to changes in class distribution. Because of this property, it is easier to see the performances of models trained by dataset oversampled by different algorithms. The goal in ROC space is to let curves be as close to upper-left-hand corner as possible, in which case the ratio between benefit and cost is maximized. To compare all test results in a more straightforward way, we also compute area under an ROC curve (AUC) which reduce the ROC performance to a single scalar value representing expected performance of the ROC curve.

5.3.5 Results. This section presents the performance of CGMOS and all the other methods on 30 real-world datasets. The same experiment procedure as the one in the experiments of the artificial dataset was conducted. All results are averged from 10 rounds of 10-folds cross-validations. A summary of the experiment results is shown in Table **??** and ROC graphs are shown in Figure **??**.

Considering the classification results of the minority class, it can be observed that the proposed approach outperforms most of the compared methods under all classification algorithms in terms of $F\text{-}score$ and $G\text{-}score$. For $F\text{-}score$ and $G\text{-}score$ of the majority class, the proposed approach in most cases is only second to the original data without oversampling. This is because the original dataset is imbalanced and it favors the majority class more than the minority class during classification. Overall, CGMOS achieves the best AUC over all tests. This is because the proposed approach takes into account both of the majority and minority classes and increases the certainties of the two classes while oversampling.

The same conclusion can be made from the ROC curves shown in Fig. **??**. It could be seen from the ROC curves that the proposed approach has the highest values almost everywhere. The proposed approach achieves the best result when random forest is used as the classifier. For b-kde as the classifier, the proposed approach gets the largest improvement since the design of the proposed approach uses b-kde for certainty computations.

To get a closer view of the performances of all compared methods on each dataset, we show the AUC results of CGMOS and all other compared methods for each dataset in Table **??**. The table shows that by average the AUC of CGMOS is 2 percent higher than SMOTE whose AUC is 2nd highest.

Previous studies show that it is not necessary for a learning procedure to obtain best classification results when a dataset is perfectly balanced[**?**][**?**]. How much to oversample is usually empirically determined [**?**]. To evaluate this aspect we performed another experiment in which we synthesized increasing number of minority samples and investigated how different amounts of new samples impact classification results.

Let δ denote the difference of data samples between the majority and the minority class. We performed multiple experiments where in each round we synthesized $k\delta$ new samples of the minority class where k gradually increased from 0.5 to 5. The classification results are shown in Figure **??**. As can be observed in the results, CGMOS achieves the best results in all cases. Also, observe that when increasing the number of data samples added, the results of CGMOS are much more robust compared with other approaches. Note that the results of some methods such as Dup(b-kde), B-SMOTE(knn) and B-SMOTE(Adaboost.M1) are even lower than the results at the starting point where datasets are not oversampled. This highlights the advantage of CGMOS when handling oversampling on boundary samples.

5.3.6 Statistical Significance Analysis. We evaluate the statistical significance of the

classification results of all competitors. Statistical significance plays a critical role in determining whether a null hypothesis should be rejected or retained, where the term null hypothesis refers to a general statement that sample observations result purely from chance. For a null hypothesis to be rejected as false, the result has to be identified as being statistically significant.

To determine whether to reject a null hypothesis, a p-value has to be calculated, which is the probability of observing an effect given that the null hypothesis is true [**?**]. The null hypothesis is rejected if p-value is less than the significance level. The significance level is the probability of rejecting the null hypothesis given that it is true. The lower the significance level the more confident we can be in replicating the results and usually the significance level is set at 5%. Then a sample observation is determined to be statistically significant if p-value is less than 5%, which is formally written as $p < 0.05$ [**?**].

We follow the same protocols used in [**?**][**?**][**?**] and choose to use Wilcoxon signed-ranks test in this paper. Wilcoxon signed-ranks test is a nonparametric statistical procedure for comparing two samples that are paired, or related [**?**]. Different from t-test [**?**][**?**][**?**] whose null hypothesis is that the mean difference between pairs is zero, the null hypothesis of Wilcoxon signed-ranks test is that the median difference between pairs of observations is zero.

The test results are shown in Table **??**. It could be seen from the table that the p-value of all tests are smaller than 0.05 and pass the test.

5.4 Conclusion In this paper, we address the imbalanced binary classification problem by proposing a novel minority oversampling strategy. Different from existing approaches, CGMOS does not randomly synthesize new data along decision boundaries. Instead, CGMOS computes the Bayes classification certainties for both the majority and minority classes and then synthesize new samples based on improvement of the certainties for samples in both

classes. We prove that CGMOS can achieve better classification results compared with SMOTE. In addition, experimental results show that CGMOS outperforms known over-sampling techniques using various metrics.

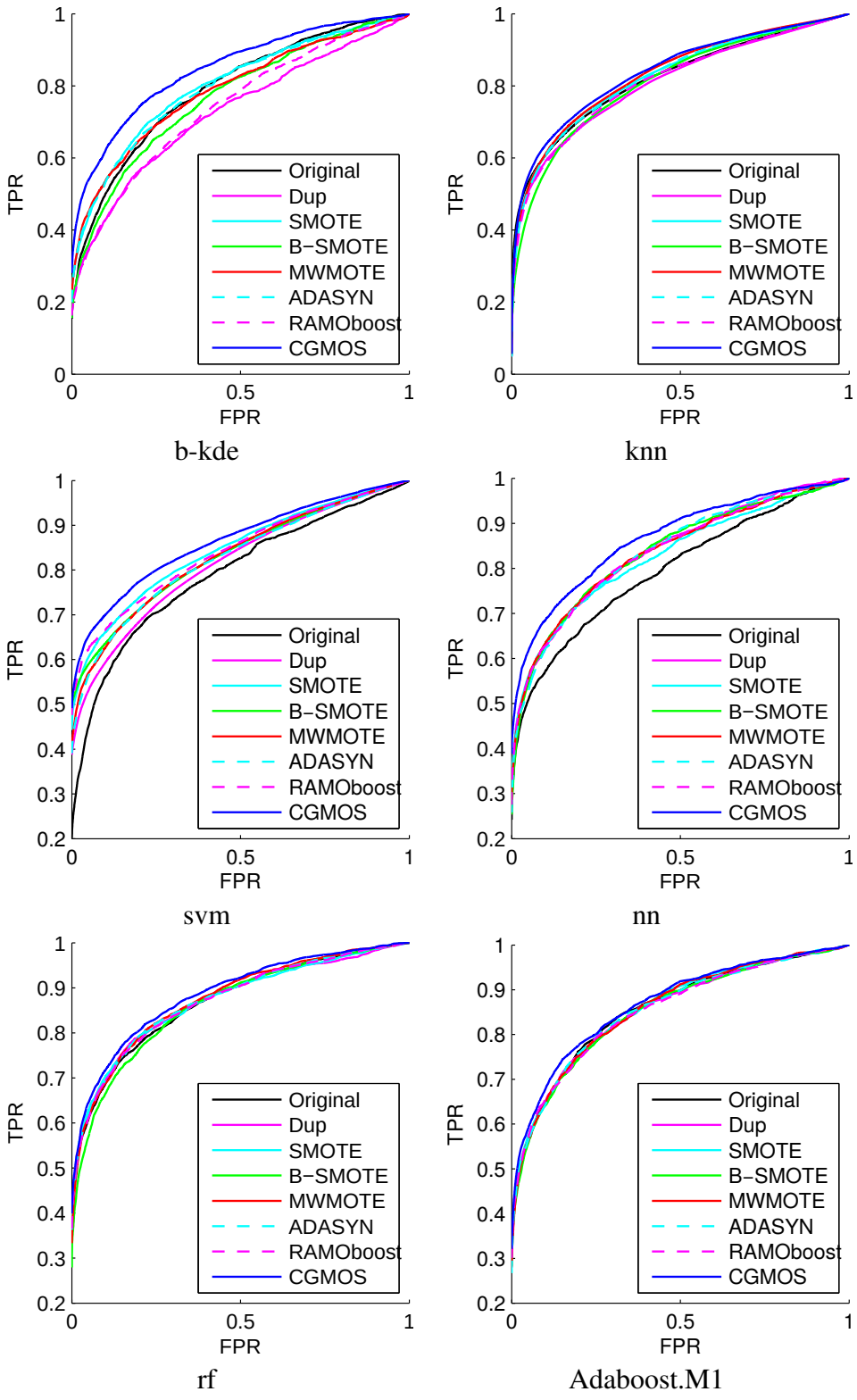


Figure 5.2. ROC curves of classification results. From left to right, up to down, we show the results of 6 different classifiers: b-kde, knn, svm, nn, rf and Adaboost.M1. Curves in blue are the results of the proposed CGMOS.

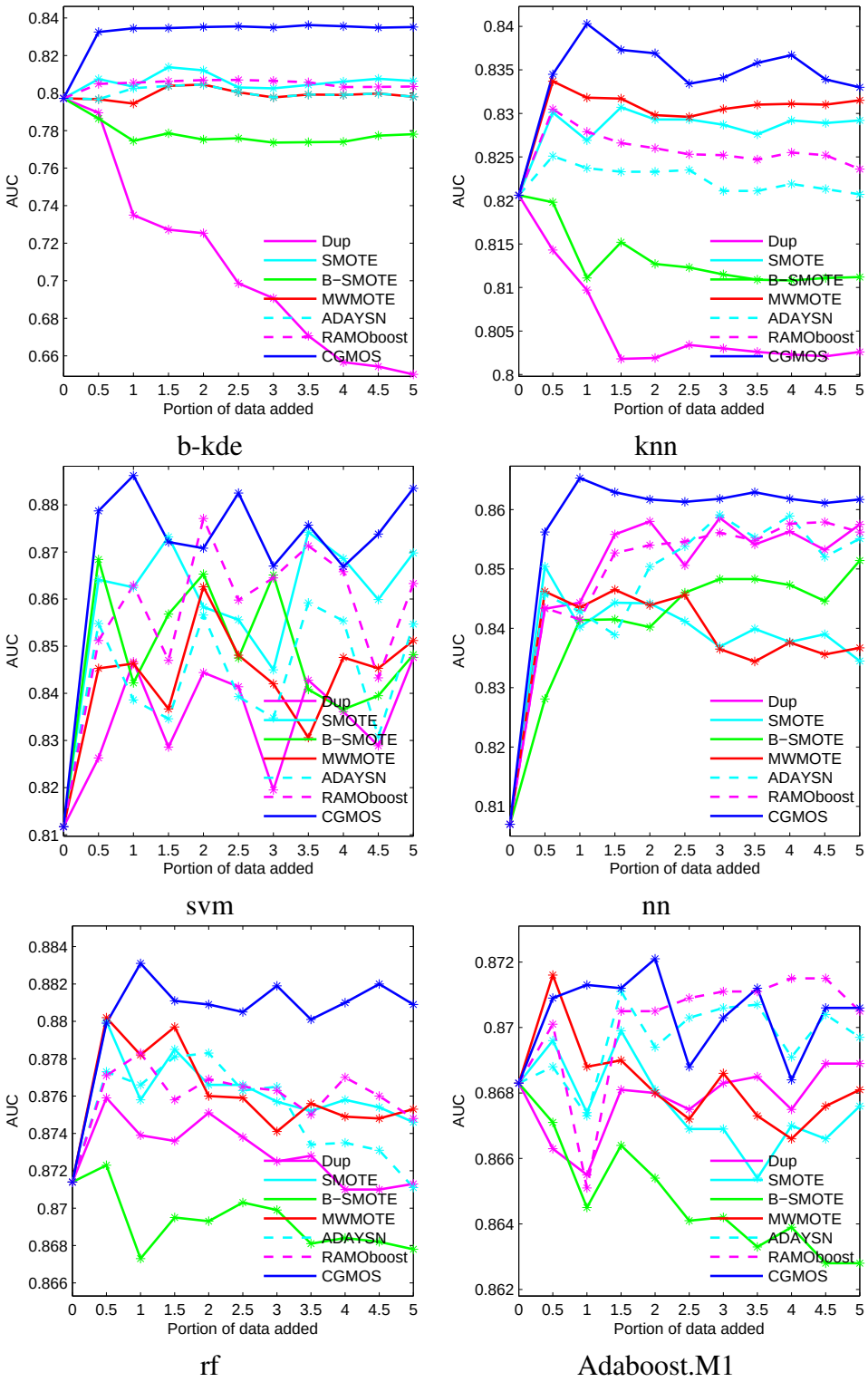


Figure 5.3. Comparison of results when increasing the number of data synthesized for the minority class. The curves measure the average AUC of the ROC curves. Curves in blue are the results of the proposed CGMOS.

Table 5.3. A summary of AUC of 8 oversampling algorithms over all 30 datasets used in our evaluation. The AUC is averaged over all 6 base classifiers used in the evaluation. It could be seen from above table that CGMOS achieves best AUC measures for 24 datasets out of 30. By average, the AUC of CGMOS is at least 2 percent higher than all other competitors.

	CGMOS	Original	Dup	SMOTE	B-SMOTE	MWMOTE	ADASYN	RAMOboost
BankMarket	**0.728**	0.661	0.708	0.718	0.710	0.721	0.710	0.723
BloodService	**0.733**	0.653	0.648	0.649	0.651	0.720	0.714	0.728
BreastCancer	0.992	0.992	**0.993**	0.992	0.989	0.991	0.991	0.992
BreastTissue	**0.984**	0.899	0.946	0.932	0.917	0.937	0.908	0.943
CarEvaluation	**0.997**	0.995	0.845	**0.997**	0.994	0.996	**0.997**	0.995
Card'graphy	**0.977**	0.976	0.939	0.962	0.956	0.925	0.957	0.960
CharacterTraj	0.985	0.962	0.717	0.985	0.978	0.981	**0.988**	0.909
Chess	**0.977**	0.974	0.959	0.973	**0.977**	0.974	0.975	0.959
ClimateSim	**0.908**	**0.908**	0.861	0.902	0.863	0.901	0.901	0.882
Contraceptive	**0.724**	0.705	0.699	0.712	0.702	0.705	0.702	0.705
Fertility	**0.673**	0.615	0.594	0.634	0.592	0.604	0.639	0.638
Haberman	**0.651**	0.623	0.577	0.600	0.593	0.594	0.587	0.586
ILPD	0.707	0.687	0.693	**0.715**	0.703	0.702	0.693	0.703
ImgSeg	**0.999**	0.998	**0.999**	0.997	0.998	0.998	0.997	0.998
Leaf	**0.908**	0.880	0.782	0.852	0.775	0.836	0.839	0.821
Libras	**0.945**	0.922	0.859	0.929	0.886	0.936	0.923	0.883
MultipleFs	**0.998**	**0.998**	0.997	**0.998**	0.997	0.997	0.996	0.997
Parkinson	0.841	0.676	0.692	0.834	0.791	0.837	**0.842**	0.760
PlanRelax	0.472	0.457	0.494	0.469	0.445	0.467	**0.488**	0.464
QSAR	**0.901**	0.886	0.879	0.895	0.863	0.886	0.886	0.882
SPECT	**0.820**	0.772	0.803	0.808	0.811	0.752	0.801	0.799
SPECTF	0.819	0.819	0.800	0.805	0.816	0.812	**0.825**	0.795
SeismicBumps	**0.743**	0.735	0.712	0.727	0.740	0.732	0.715	0.691
Statlog	**0.998**	0.992	0.996	**0.998**	0.990	0.996	0.976	0.996
PlatesFaults	**0.956**	0.928	0.844	0.954	0.920	0.943	**0.956**	0.881
TAEvaluation	**0.748**	0.682	0.644	0.703	0.671	0.707	0.665	0.657
UserKnowledge	**0.958**	0.837	0.919	0.953	0.947	0.951	0.950	0.888
Vertebral	**0.890**	0.839	0.869	0.855	0.829	0.860	0.794	0.872
Customers	**0.952**	0.930	0.943	0.946	0.884	0.902	0.946	**0.952**
Yeast	**0.925**	0.792	0.844	0.907	0.898	0.900	0.906	0.851
Average	**0.864**	0.827	0.808	0.844	0.830	0.842	0.842	0.830

Table 5.4. A summary of p-values of statistical significant tests of classification results using CGMOS against each of all the other competitors.

	Knn	Rf	B-kde	Nn	Svm	Boost
Original	5e-5	1e-4	0.004	1e-4	0.026	0.04
Dup	2e-6	5e-5	3e-6	0.03	0.049	0.004
SMOTE	0.003	2e-4	6e-6	0.018	0.006	0.046
B-SMOTE	4e-6	7e-6	2e-5	5e-4	0.047	5e-4
MWMOTE	0.046	4e-5	1e-5	0.003	0.005	0.007
ADASYN	8e-6	7e-5	9e-5	0.005	1e-4	0.003
RAMOboost	2e-6	5e-5	3e-6	0.001	0.045	0.035

CHAPTER 6

CONCLUSION

Synthetic data is extensively used in computer vision area. Although the success of using synthetic data has been shown by many work in the past decades. Usage of synthetic data in more general problems is limited by several factors. First, most work creates and uses synthetic data in a application-specific way that it is hard for other application to benefit. It is interesting if there is an efficient and more generous way to create and use synthetic data, so the method is eligible to most of researches and applications. Second, improvement gained by using synthetic data is limited by a *synthetic gap* existing between actual data and synthetic data. To even boost the performance and unleash the power of synthetic data, such gap should be resolved from many aspects including feature design, learning algorithm design, the way of eliminating the gap and so.

Towards these problems and challenges of using synthetic data. My research has been primarily focusing on solving computer vision related classification and recognition problem using synthetic data. Ways of creating and using synthetic data have been proposed and proved to be effective in my research. To further improve the performance that could be brought by synthetic data, I study the problem of *synthetic gap* and propose many ways to overcome such inherent difference between actual data and synthetic data. To sum up, my doctoral research have following primary contributions.

First, in most of learning based vision research, training data is very valuable and hard to collect, which usually lead to an unsatisfactory machine learning result. Conventional data augmentation method only creates new data which has only small variance from existing ones. In this way, data augmentation fails to introduce many new informations about data and doesn't increase enough data coverage. Thus the performance brought by data augmentation is constraint. In my research, to better simulate actual data, a series of methods that create synthetic data are introduced. A significant problem that data argu-

mentation does not solve and I have to figure out is what makes real data looks like real data and how to simulate unseen data using existing ones. I believe transformation in either data space or parameter space among existing data is the key to this problem. Therefore, methods with new techniques such as: data congealing, registration, parametrization are invented. Different from data argumentation which applies transformation based on single data only, my proposed method learns transformation among data, thus my methods are able to generate synthetic data with larger variance if projected to parameter and feature space which leads to a better simulation of unseen actual data.

Second, directly use synthetic data will most likely cause a failure or unsatisfactory performance in machine learning algorithm. That is mostly because the inherent difference between actual data and synthetic data created using a parametric prototype, which is defined as a *synthetic gap* in my work. In my researches, I eliminate the impact of such difference in two different phases of a machine learning procedure which are in feature extraction and in learning algorithm design. In feature extraction, I designed features which are able to capture the common characteristics of both actual and synthetic data, while ignoring inherent pattern and noises of actual data. In learning algorithm design, learning algorithms are designed to be able to deal with the difference between two kinds of data.

Third, to eliminate the *synthetic gap*, a general framework based on a neural network with novel architecture is proposed. The neural network is trained as a regressor to automatically find out the relationship between two types of data, thus is able to close the gap between two data. With this mechanism synthetic data could better simulate actual data and used in machine learning procedure.

Fourth, create synthetic data in data space and compute sophisticated features from the data is a complicated task. Creating synthetic data in data space narrow down the range of using synthetic data in learning, not only because it requires enough pre-knowledges of problems and data, but also for most problems there is almost no way a synthetic data can

be built in data space to mimic the actual ones. A better solution but much more challenging is to directly create synthetic data in feature space. Inserting new data in feature space is hard. Because data representation in feature space is abstractive that the actual data space representation of a location in feature space is unknown or hard to show. The problem I need to solve here is to figure out answers of two questions: where I are going to insert a new data and why I insert it there? In my research, I propose a general algorithm that create new synthetic data through learning the spatial relation of existing data in feature space. The algorithm first learns distribution and structure from existing data, then each new data is created to preserve these knowledges.