Active Learning for Sketch Recognition
and
Active Scene Learning

by

Erelcan Yanık

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Engineering

Koç University

2 September, 2013

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Erelcan Yanık

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assist. Prof. T. Metin Sezgin

_____

Assoc. Prof. Yücel Yemez

_____

Assoc. Prof. Engin Erzin

Date: _____

ABSTRACT

Sketching is a natural and effective means for expressing and sharing ideas. These qualities have made sketching an emerging interaction modality in pen-based systems. Sketch-based interfaces rely on the availability of accurate sketch recognition engines, which in turn require large amounts of labeled data for training. Unfortunately, labeling sketch data is time consuming and expensive, because it requires the involvement of human annotators. We demonstrate the utility of the active learning technology in reducing the amount of manual annotation required to achieve target recognition accuracy.

The first part of our work presents the first comprehensive study on the use of active learning for isolated sketch recognition. We present results from an extensive analysis which shows that the utility of active learning depends on a number of practical factors that require careful consideration. These factors include the choices of batch selection strategies, informativeness measures, seed set size, and domain-specific factors such as feature representation and the choice of database. Since active learning community lacks such factor based analysis, our empirical analysis is examplary. Our results imply that the Margin-based informativeness measure consistently outperforms other measures. We also show that the use of active learning brings definitive advantages in challenging databases when accompanied with powerful feature representations.

The second part of our work deals with active learning on sketches containing more than one object, the so-called "scenes". We propose a new active learning framework which distinguishes itself from the previous frameworks in terms of the segmentation aspect, properties of the training instances and multi-class labeling capability for the scenes. We propose two selection schemes under this framework: scene-wise selection

and segment-wise selection. We adapted several basic methods of previous frameworks to our scene-wise selection scheme. We show that selection schemes we proposed under our framework outperform random selection. Finally, we show that our segment-wise selection outperforms the methods following scene-wise selection scheme.

ÖZETÇE

Çizim, fikirlerin ifade edilmesi ve paylaşılması için doğal ve etkili bir araçtır. Bu nitelikler çizimin kalem tabanlı sistemler için yeni bir etkileşim kipi olmasını sağlıyor. Çizim tabanlı arayüzlerin kullanılabiliriliği başarılı çizim tanıma sistemlerinin varlığına dayanır ki bu da çok sayıda etiketlenmiş verinin model eğitimi için kullanılmasını gerektirir. Ne yazık ki, çizim verisini etiketlemek zaman alıcı ve masraflıdır çünkü etiketleme için insanların katılımı gereklidir. Bu çalışmada, hedeflenen tanıma başarısı için gerekli manuel etiketleme yükünün azaltılmasında "aktif öğrenme" teknolojisinin kullanılabilirliğini gösteriyoruz.

Çalışmalarımızın ilk bölümü, tekil çizimlerin tanınması için aktif öğrenmenin kullanımına ilişkin yapılmış ilk kapsamlı empirik analizi içermektedir. Analiz sonuçlarımız göstermektedir ki aktif öğrenmenin kullanılabilirliği bir dizi faktörün dikkatle göz önünde bulundurulmasını gerektirmektedir. Bunlar "batch selection" stratejileri, "informativeness" ölçütleri, başlangıç seti büyüklüğü; öznitelik vektörü ve veritabanı seçimi gibi alana spesifik faktörlerdir. Empirik analizimiz, faktör bazlı analizden yoksun olan aktif öğrenme literatürü için örnek teşkil etmektedir. Analiz sonuçlarımız "Margin based informativeness" ölçütünün diğer "informativeness" ölçütlerinden tutarlı olarak daha üstün performans sergilediğini göstermektedir. Buna ek olarak aktif öğrenmenin zorlu veritabanlarında güçlü öznitelik vektörleri eşliğinde daha verimli olduğunu gösteriyoruz.

Çalışmalarımızın ikinci bölümü, birden fazla nesne içeren sahne olarak adlandırdığımız çizimleri tanıma için aktif öğrenme kullanımını sağlamak üzerinedir. Yeni bir aktif öğrenme sistemi ortaya koyuyoruz; öyle ki bu sistem bölükleme yaklaşımı, model eğitimi için kullanılan örneklerin özellikleri ve sahneler için çoklu-sınıf etiketlenebilirliği bakımından önceki sistemlerden ayrılmaktadır. Bu sistem altında da iki ana "örnek

seçim" stratejisi ortaya koyuyoruz ki bunları sahne bazında ve segment bazında seçim olarak adlandırıyoruz. Önceki sistemlerdeki bazı temel yöntemleri sahne bazlı seçim stratejimize adapte ettik. Önerdiğimiz hem sahne bazlı hem de segment bazlı örnek seçim stratejileri rastgele örnek seçme stratejisine üstünlük sağlamıştır. Buna ek olarak, segment bazlı örnek seçim stratejisinin sahne bazlı örnek seçim stratejisinden daha iyi performans gösterdiğini de gösteriyoruz.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor T. Metin Sezgin. His guidance was encouring and essential throughout my years at grad school. He was always available and resourceful.

I was fortunate to be part of the IUI Laboratory. I would like to thank Atakan Arasan, Ayşe Küçükyılmaz, Banuçiçek Gürcüoğlu, Burak Özen, Çağla Çığ, Özem Kalay and Sinan Tümen. They were all inspiring colleagues.

I was lucky to have exceptional friends. I would like to thank Alhun Aydın and Burak Özaydın. They always kept my mind alive and excited via our discussion meetings.

I would like to thank my friend of eleven years, Şeref Ayyavuz. Without his encouragement and support, I could have given up many challenges.

I would like to thank my mother Emel Yanık, my father Erdal Yanık, and my brother Emrecan Yanık. My family was the primary source of emotional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

x, y	Input data and corresponding label
$I(.)$	Informativeness score
$H(.)$	Entropy
$D(. \parallel .)$	Kullback-Leibler divergence

Chapter 1

INTRODUCTION

1.1 What is Sketch Recognition?

Sketching is a natural and effective means for expressing and sharing ideas. These qualities have made sketching an emerging interaction modality in pen-based systems. By sketch, we mean messy, informal hand-done drawings. Recognizing sketches with objects of a symbolic nature is our main concern.

We can group sketches under two main titles: isolated sketches and scenes. Isolated sketches consist of single object (symbol) whereas scenes can contain more than one object. Also we can categorize scenes as "interspersed" or "non-interspersed", depending on the way they are drawn. Non-interspersed sketches impose a constraint on the drawing style such that a user must complete drawing an object before starting a new object in a scene. If user starts drawing a new object before completing the current one, the sketch becomes interspersed. When the strokes of the sketch are ordered by the time they were drawn, an object may not have all of its strokes lined in a consecutive manner.

Recognizing an isolated sketch means identifying the object (symbol) and assigning a class label. To recognize a scene, we need to understand which part of the sketch constitutes which objects. Therefore, localizing objects in a scene is required before (or simultaneously with) assigning class labels to the objects in the scene. Recognizing isolated sketches is straightforward. A model trained with the samples of each object class is sufficient for isolated sketch recognition. However recognizing scenes introduces the segmentation problem. Segmentation is the task of grouping strokes

(or fragments) so that those constituting the same object end up in the same group. Note that it is not known what object the strokes (or fragments) form.

1.2 What is Active Learning?

Active learning is a subfield of machine learning. It allows supervised machine learning algorithms to ask questions about data. Moreover, a learning algorithm can choose the data from which it learns when active learning is employed.

Supervised machine learning algorithms require labeled data. If existing data is unlabeled, users (supervisors) tend to label all the data (or subset of the data) at once. However labeling an instance might be very costly for some systems. For example, visually selecting which image segments belong to an indicated object (on SIVAL repository) takes time from 60 seconds up to 204 seconds [1]. Even more effort consuming, annotating an article (from CKB database) with only four entities (actor, organization, role, location) takes time from 56 seconds up to 1000 seconds. When excess amount of such data is available, labeling all the data is infeasible. Rather than labeling all the data, labeling randomly selected subset of the data is an option. However, this may cause labeling unnecessary data. However selecting instances wisely by employing active learning will reduce labeling effort.

Active learning is a machine learning strategy that aims to reduce the labeling effort by selecting the most informative samples from a pool of unlabeled data. Active learning process is initialized by training the model with a few labeled samples, the so-called "seed set". The process continues in rounds until available resources are consumed or a target validation accuracy is reached. A round starts with classifying the unlabeled samples in the pool with the existing model. By using these classification results, informativeness of each sample is measured in order to represent how useful a sample is. Then, a batch of samples are selected by avoiding redundant information, and user is asked to label these samples. The round ends with adding the newly labeled data to the training set and re-training the model.

1.3 Why do Active Learning for Sketch Recognition?

Sketch recognition can be cast as a supervised machine learning problem. Since we need labeled data to solve this problem, active learning is a desirable option to integrate into data labeling process. However there is no active learning attempt for sketch recognition to our knowledge, in the literature. One reason is that researchers/developers tend to collect data in a strictly task-specific manner such that data is collected as isolated sketches of which the data provider is asked to draw an object with the given name. For example, a developer may collect sketches of a car from the data provider which are labeled automatically as "car". Hence, re-usability of the data is not considered in such cases. However finer/detailed labels can be given by the developer (or the other developers) aftwerwards, for enabling a large variety of tasks by employing active learning.

One other reason might be not allowing dynamic user-integrated systems since developers generally collect data and build the systems to its final version. Hence, end-users are only allowed to use or manipulate the system only over the pre-defined set of object classes. Constraining the capability of a system on learning new set of object classes or defining new tasks is one of the major drawbacks in sketch recognition. An intelligent system should allow users to define new object classes and new tasks especially when excess amount of data is avaialble to the users. For example, consider a country-wide educational application which can review free-form answers of exams (quizes, homeworks etc.) and give feedback to the instructors and the students. These feedbacks might include examples of correct answers (solved in various ways), common mistakes, suggestions on solutions and so on. To design such a system, the developer must allow the instructors (or the system manager) to create new questions which require defining object classes and tasks (e.g. solutions and content of the feedback). At this point, active learning will assist instructors to achieve their desired model with labeling very few data rather than labeling all the data coming from students. As in this example, enabling end-users to use their own data to train such a dynamic system will result in a large number of user-generated applications. Since there will

be existing data (e.g. drawn data for sketch recognition) already, support of active learning will be crucial.

Active learning encourages the researchers/developers to be attentive on the reusability of data and pave the way for a variety of sophisticated sketch-based systems. Moreover, active learning advancements for sketch recognition will be applicable in many areas. For instance, state-of-the-art solution for segmentation of non-interspersed scenes is a 1D dynamic programming based solution. This is indeed the optimal solution for segmentation of any sequential data. For example, we can segment and recognize a series of gestures/activities of humans and robots by employing the same solution. Isolated instances of such problems might occur in seconds, even in milliseconds. Hence, labeling such sequential data is costly in terms of both effort and time spent. Therefore, developments in active learning for sketch recognition is applicable for reducing labeling efforts in areas utilizing sequential data. Indeed, as long as the definition of segmentation holds these developments can be used in areas utilizing non-sequential data as well.

1.4 Thesis Statement

This thesis aims to investigate the key aspects of active learning for sketch recognition. We present the first extensive empirical analysis on basic active learning methods for sketch recognition. This analysis targets to examine the factors affecting the active learning performance and presents valuable results and insights for how to make a better use of active learning. We aim our factor based analysis serves as an example for active learning community for conducting comprehensive analysis. In order to enable active learning on scenes, we propose a new active learning framework which distinguishes itself from the previous frameworks in terms of the segmentation aspect, properties of the training instances and multi-class labeling capability for the scenes. We claim and show that selection schemes we proposed under this framework outperforms random selection. Finally, we claim and show that segment-wise selection scheme outperforms scene-wise selection scheme.

1.5 Thesis Roadmap

Chapter two presents our empirical analysis on active learning methods. Chapter three introduces our new active learning framework and the selection schemes we proposed under this framework. Chapter three also provides a performance analysis for these selection schemes. Both chapter two and chapter three contain sections presenting experimental design, analysis methodology, results and discussion. Chapter four provides the related work for our work presented in chapter two and chapter three. Chapter five summarizes future work. We conclude by summarizing our contributions and provide a general discussion.

Chapter 2

EMPIRICAL ANALYSIS OF ACTIVE LEARNING METHODS

2.1 Introduction

The literature has a variety of theoretically well-defined active learning methods. However, it is an open question which one to use on a new database. There are methods with empirical justification which might act as a heuristic guide in some domains such as sequence labeling tasks, image segmentation, image retrivial and image categorization. However there is neither an active learning attempt nor empirical analysis on active learning in the literature for sketch recognition, to our knowledge.

Despite its theoretical appeal, recent empirical results show that active learning does not always yield the expected benefits in practical real world problem settings [2]. For example, Schein and Ungar report inconsistent and negative results for active learning methods [3]. Likewise, Gasperin reports that none of the experimented active learning methods reached a remarkable performance although they select different sets of samples from each other [4]. Guo and Schuurmans also point out that active learning methods they employed performed poorly with respect to random learning which is the strategy of selecting unlabeled samples randomly from the pool [5]. Hence, there is a practial and real need for analysing the empirical performance of active learning in various settings.

Existing empirical analyses lack of questioning the factors affecting the active learning performance. They rather compare a few methods with their proposed methods and report results in terms of only the performance superiority. Since literature points out that active learning might yield unsatisfactory results, this implies that

active learning performance is not only affected by the employed method but also there should be other factors in effect. Therefore, detecting and investigating such factors will play a crucial role to improve the active learning field.

We investigate the performance of the combinations of a large variety of informativeness measures and batch selection strategies under factors such as database, seed set size and feature representation, for sketch recognition. Our analysis results constitute a detailed and practical guide for active learning users for sketch recognition. Not only our results but also our experimental design will direct researchers to a better use and analysis of active learning methods.

This chapter is organized as follows: First, we introduce informativeness measures and batch selection strategies that are included in our analysis. In Section 3, we first describe the databases and the feature representations used in our experiments, then describe the details of our experimental design. In Section 4, we describe the deficiency metric employed in our analysis and then present the analysis methodology. We present the analysis results in Section 5. We conclude the chapter with a disscussion.

2.2 Active Learning Methods

There are two essentials in the process of active learning: measuring informativeness of unlabeled samples and avoiding redundancy when selecting a batch of informative samples. In this section, we describe informativeness measures and batch selection strategies that are used in our experiments.

2.2.1 Informativeness Measures

There are two main approaches to measure informativeness: the single classifier approach and the query by committee (QBC) approach. The rationale is that samples that a classifier cannot confidently classify, or a group of classifiers disagree on have the potential of supplying more information when labeled. We briefly describe the informativeness measures included in our experiments in Table 2.1. Now, we will explain informativeness measure in more detail.

Table 2.1: Brief description of the informativeness measures included in our analysis

	Informativeness Measures	A sample is informative when:
Single Classifer Approach	Entropy based Selection [6, 7] *	The entropy is high on class probabilities of a sample.
	Least Confident based Selection [6, 8]	The most likely class probability of a sample has a low value.
	Margin based Selection [6] *	The difference of the most and the second most likely class probabilities of a sample has a low value.
	Körner-Wrobel Selection [9] *	The Körner-Wrobel value computed for the sample is low. It is a combination of the maximal class probability and the margin based informativeness.
Query by Committee Approach	Kullback-Leibler Divergence based Selection [9, 10, 11]	KL-Divergence among the committee on a sample is high.
	Jensen-Shannon Divergence based Selection [9, 12]	JS-Divergence among the committee on a sample is high.
	Vote Entropy based Selection [9, 13]	The entropy of the class label votes of the committee is high.
	Weighted Vote Entropy based Selection [9]	The weighted entropy of the class label votes of the committee is high.

* The method has implementation also for the query by committee approach, in the literature.

Single Classifier Approach

Single classifier approach queries the samples in the guidence of a single classifier. If the classifier has less confidence on classifting a sample it is more informative. In the literature, this is referred as the "confusion/uncertainty of the classifier". The intuition is that confusion of the classifier increases if an unlabeled sample is closer to the decision boundary of the model. Also the change in the decision boundary is expected to be high when such a sample is added to the training set. We experimented with four of the single classifier based informativeness measures:

Entropy based informativeness: This measure computes entropy in the probability vector containing the probabilities for each class label predicted by the current model for a sample. This value is assigned as the informativeness of that sample. The higher the value of the entropy, more uncertain the classifier is on that sample. Hence, user is requested to label the sample with the highest entropy value. Entropy based informativeness score can be formalized as follows:

$$I_{Entropy}(x) = -\sum_{i} P(y_i|x) \log P(y_i|x)$$

where $P(y_i|x)$ is the probability of unlabeled sample x belonging to class y_i, predicted by the current model. Also y_i ranges over all possible class labels.

Least Confident based informativeness: This measure computes the confidence of the classifier on an unlabled sample by using the probability of the class label which is assigned as the most likely class label by the classifier. If this probability is low, then the classifier is not confident of its prediction on that sample. Least confident based informativeness can be computed as:

$$I_{LC}(x) = 1 - P(y^*|x)$$

where $P(y^*|x)$ is the probability of the class label assigned to the most likely class label (y^* indicates the most likely class label).

Margin based informativeness: This measure computes the difference in the predicted probabilities for the most likely class label and the second most likely class label. Therefore, this measure considers not only the confidence on the predicted most likely class label but also how much more the classifier is confident on the most likely prediction over the second most likely prediction. Hence, as the difference between the most likely two labeling decreases, informativeness of sample increases. Margin based informativeness can be computed as:

$$I_{Margin}(x) = -(P(y_1^*|x) - P(y_2^*|x))$$

where $P(y_1^*|x)$ is the probability assigned to the most likely class label and $P(y_2^*|x)$ is the probability assigned to the second most likely class label(y_1^* indicates the most likely class label and y_2^* indicates the most likely class label).

Körner-Wrobel informativeness: This measure combines the probability of most likely class label, the margin based informativeness and the number of classes. The lower the value of Körner-Wrobel informativeness is, the more informative the sample is. We can formalize it as:

$$I_{K-W}(x) = I_{Margin}(x) + 0.5\frac{1}{(|C|P(y^*|x))^3}$$

where $|C|$ is the number of the classes and $P(y^*|x)$ is the probability of the class label assigned to the most likely class label (y^* indicates the most likely class label).

Query by Committee Approach

Query by committee approach queries the samples in the guidence of a committee of classifiers. If the committee members have high disagreement on how to label an unlabeled sample, that sample is expected to be very informative. The aim of QBC approach is to minimize the version space which is the all hypothesis consistent with the labeled data. In other words, version space can be thought as a set of all possible models that we can generate. Imagine the committee members as investigators searching a set of models in order to find the best model. The aim is to constrain the version space as much as possible by eliminating the controversial parts. We experimented with four of the query by committee based informativeness measures:

Kullback-Leibler Divergence based informativeness: The Kullback-Leibler divergence is a non-negative measure of the divergence between two probability distributions p and q in the same event space $X = \{x_1, ..., x_c\}$ and computed as:

$$D(p \parallel q) = \sum_{i=1}^{c} p(x_i) log \frac{p(x_i)}{q(x_i)}$$

Kullback Leibler Divergence to the mean is assigned as KL Divergence based informativeness. It is the average KL divergence between the each distribution and the mean of all distributions. Therefore, KL Divergence based informativeness can be computed as:

$$I_{KL-Div}(x) = \frac{1}{k} \sum_{i=1}^{k} D(p_i(x) \parallel p_{mean}(x))$$

where k is the number of classifiers in the committee and $p_i(x)$ is the probability distribution for unlabeled sample x given by the i^{th} classifier. As the value of KL Divergence to the mean is high for a sample, it means that the committee has a high disagreement on that sample and it is expected to be very informative.

Jensen-Shannon Divergence based informativeness: This measure computes the Jensen-Shannon Divergence for k distributions:

$$I_{JS-Div}(x) = (p_1(x), ..., p_k(x)) = H(\sum_{i=1}^{k} w_i p_i(x)) - \sum_{i=1}^{k} w_i H(p_i(x))$$

where p_i is the class probability distribution given by the i^th classifier for sample x, w_i is the weight of the i^{th} classifier in the committee, and $H(p_i(x))$ is the entropy of the class probability distribution given by the i^{th} classifier. We determine weights by considering accuracies of the classifiers on the labeled data, in our experiments. A high value of Jensen-Shannon Divergence increases indicates that the disagreement of the committee members is high and the sample is informative.

Vote Entropy based informativeness: This measure is computed by considering the votes given by committee members over the class labels for a sample. We can compute it as:

$$I_{VE}(x) = -\frac{1}{\log k} \sum_{i=1}^{|l|} \frac{V(l_i, x)}{k} \log \frac{V(l_i, x)}{k}$$

where k is the number of classifiers in the committee and $V(l_i, x)$ is the number of classifiers assigning label l_i to the sample x. If value of vote entropy is higher for a sample, it is more informative.

Weighted Vote Entropy based informativeness: This measure is computed vote entropy. In addition, this measure considers the performances of the committee members. We can compute it as:

$$I_{WVE}(x) = -\frac{1}{\log w} \sum_{i=1}^{|l|} \frac{W(l_i, x)}{w} \log \frac{W(l_i, x)}{w}$$

where w is the sum of the weights of all classifiers in the committee and $W(l_i, x)$ is the is the sum of the weights of the committee members assigning label l_i to the sample x. If value of weighted vote entropy is higher for a sample, it is more informative. For determining weights, we consider the accuracies of the classifiers on the labeled data, in our experiments.

2.2.2 Batch Selection Strategies

Due to re-training cost at each round of active learning process, adding a batch of samples to the training set rather than single sample, at a time, is a desirable option. Although this idea reduces time and computational power requirements, it brings

the problem of selecting samples with redundant information. For example, the most informative two samples might contain so similar information that adding both to the training set does not yield a performance difference than adding one of them. Hence, we should avoid selecting such samples together. To solve this problem, several batch selection strategies were proposed in the literature and we used four of them in our experiments. Our empirical analysis includes the following batch selection strategies:

Default selection: N samples with top informativeness scores are selected in the Default batch selection strategy, where N is the batch size. We employ it in our experiments as a baseline strategy.

Global-FV strategy: This strategy is based on clustering as Shen's strategy [14]. We select N samples with top informativeness scores and divide them into K clusters. Then, we choose cluster centers as the batch, where K is the batch size. We define N as R*NumberOfClasses such that we fix R to a small number as 3 in order not to depreciate the effect of informativeness measure. Global-FV uses feature vectors for clustering.

Global-PE strategy: This strategy is same as Global-FV strategy, except it handles clustering by considering class probability distribution of the samples in the pool. By such a strategy, we aim to observe whether the feature representation or the probability estimates predicted by the current classifier is favorable. Hence, rather than pre-defined feature values, we consider current confusion of the classifier in order to cluster the data in the pool.

Combined strategy: Combined strategy is implemented as in Brinker's paper [15]. In this method, informativeness scores are computed first. Then, we keep a "selected set" which we add samples to be labeled one by one. The most informative sample is added to the list as the first. Then, we compute diversity of the remaining unlabeled samples with respect to the selected list. Diversity is the maximum value of the cosine similarity between an unlabeled sample and the samples in the selected set. After computing diversity values, we compute equally weighted average of the informativeness score and diversity score for remaining unlabeled samples. The unlabeled sample

with the highest score is added to the selected set. This process continues till the size of the selected set reaches the batch size.

2.3 Experimental Design

We describe databases and feature representations used in our experiments and explain the design details of our experiments in this section.

2.3.1 Databases and Feature Representations

We conducted our experiment on two databases. The first database is our Course of Action Diagrams (COAD) database [16], and the other database is the publicly available Nicicon database [17]. The COAD database contains 620 samples and 20 different classes whereas the Nicicon database contains 22958 samples and 14 different classes.

We used two image-based methods for feature extraction: Zernike Moments [18] and IDM [19]. We set order parameter of Zernike Moments to 12. For IDM, we set its kernel size, sigma and resample interval parameters to 25, 2 and 100 respectively.

To investigate the effects of database and seed size factors; we created datasets for each combination of these factors as depicted in Figure 2.1. Type of the database, either the COAD or the Nicicon database, is selected at the beginning of the dataset creation process. Seed size denotes the number of labeled samples used to initialize an active learning (or a random learning) process. We tried two seed size values: 1 sample per class (small) and 4 samples per class (large). In order to perform a statistical analysis after conducting the experiment, we repeat the dataset creation process 10 times and obtain 10 different datasets, for each combination of database and seed size parameters. Note that there are 4 parameter combinations, hence 40 datasets in total.

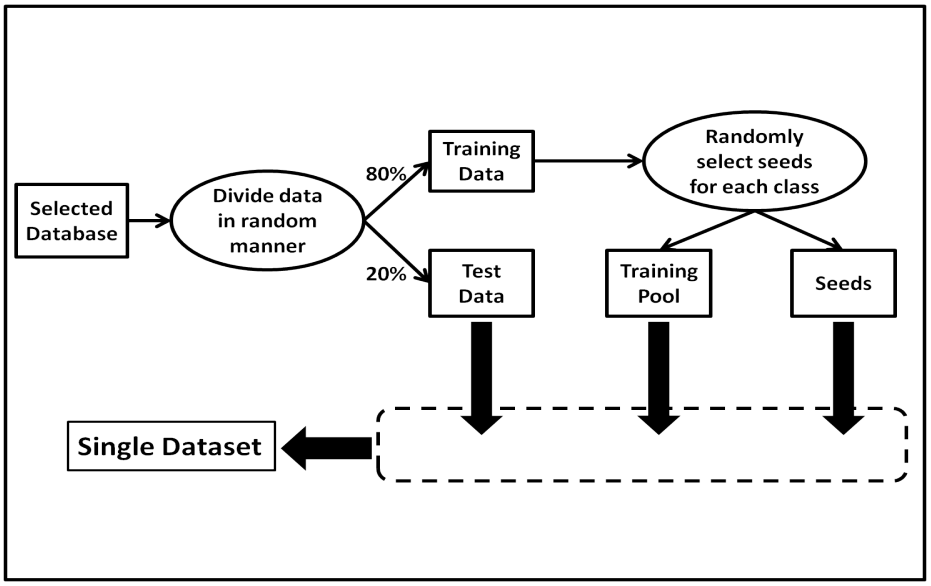


Figure 2.1: The flow chart of the process of creating a single dataset.

2.3.2 Trials

A trial refers to a process of active learning (or random learning) that we measure a classification accuracy of the model at each round. We initialized each trial by training the classifier or the committee members with the seed set. Then, the process continued by selecting and adding 10 samples to the training set and re-training the model at each round. Each trial is continued until all the unlabeled data in the training pool are labeled for the COAD database. Since the Nicicon database has a fairly large training pool, we limited its trials to 120 rounds. This limit is sufficient for the classifier accuracies to converge.

We conducted trials for all combinations of 2 feature representations, 4 batch selection strategies and 8 informativeness measures for each dataset. Also we did two more trials for random selection with 2 feature representations. Therefore, we conducted 66 trials for each dataset as demonstrated in Figure 2.2. In total, we conducted 2640 tests.

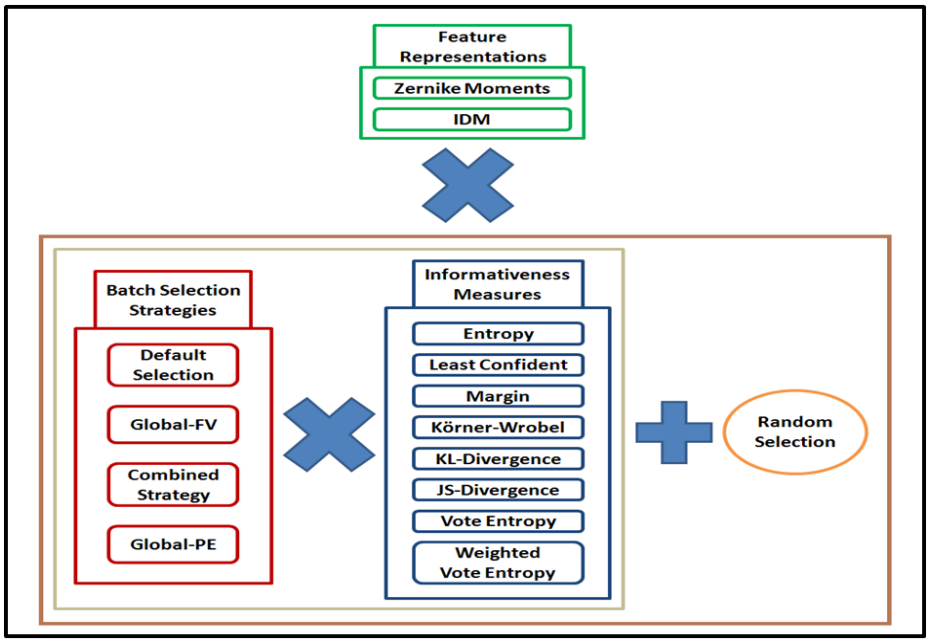


Figure 2.2: For a single dataset, trials for each combination of batch selection strategies and informativeness measures; and a trial for random selection were conducted, over two feature representations.

2.3.3 Classifier

We employed probabilistic SVM with RBF kernel both for the single classifier and the query by committee (QBC) approaches. We used 4 classifiers in the committee for QBC approach. We utilize the grid search and 5-fold cross validation to tune parameters of the SVM while re-training the model.

2.4 Analysis

2.4.1 Deficiency Metric

In order to assess the relative performance of various active learning methods, we used the deficiency metric as described by Baram et al. [20]. The deficiency metric gives a standardized measure of the relative performance of algorithms throughout the active learning process. To compute the deficiency value, accuracies by methods at each round are used as demonstrated in Figure 2.3. Areas between two accuracy curve

and the maximal accuracy line are computed first. Then, the ratio of two areas gives the deficiency where area above the reference curve is in the denominator. Deficiency values closer to zero indicate that the method in question performs better than the baseline. A value of one indicates that the methods have comparable performance whereas a deficiency value higher than one implies that the reference method performs better.

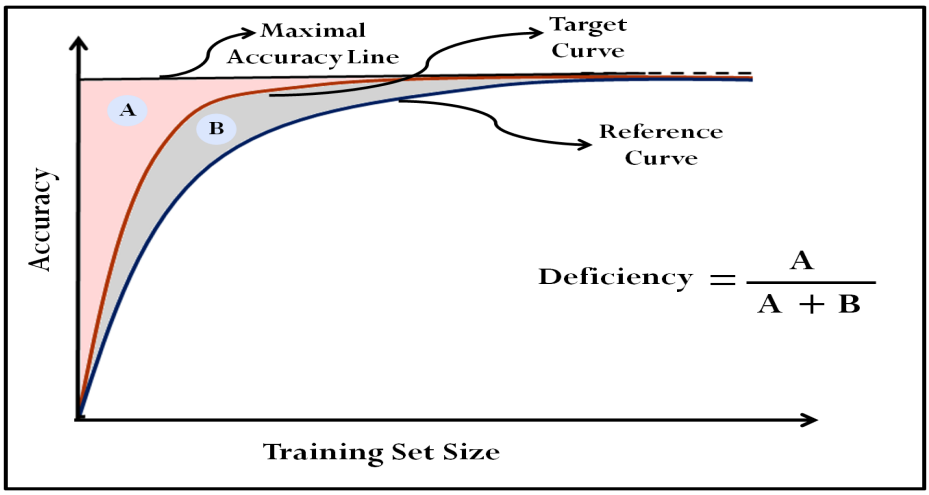


Figure 2.3: The graphical description of the deficiency metric.

2.4.2 Analysis Methodology

In order to assess the statistical significance of the differences observed in the deficiencies obtained from different active learning setups, we conducted ANOVA. Throughout our analysis, we performed Mauchy's sphericity test to check whether the variances of the differences between all possible group pairs, subject to ANOVA, are equal. If sphericity is violated, the degrees of freedom are corrected using Greenhouse-Geisser correction. We also performed Levene's test for checking homogeneity of variances between groups and used transformed values where appropriate. Bonferroni corrected paired t-tests were performed as Post-Hoc tests.

We conducted 5-way Mixed ANOVA with between group variables of database and seed size; and within group variables of feature representation, batch selection strategy

and informativeness measure. The deficiency values were taken as dependent variable which were computed for each active learner with respect to the random learner.

To further observe the effect of seed size on the performance of active learners, we conducted a 4-way Mixed ANOVA. For this design, between group variable is database; and within group variables are feature representation, batch selection strategy and informativeness measure. The deficiency values were taken as dependent variable. However, these deficiency values measure the performance gap between using the large seed set and the small seed set for active learners. Therefore, deficiency values were computed for active learning performance with the large seed set with respect to active learning performance with the small seed set.

2.5 Results

2.5.1 Choice of Informativeness Measure

Selecting an informativeness measure is the first step of an active learning user. Although this is the most crucial step, there is no certain instruction for selecting the "right" informativeness measure. However previous works on similar domains and feature representations function as heuristics to select a working informativeness measure. In this subsection, we will present our analysis results on the informativeness measure factor and show that the Margin based informativeness measure is desirable on sketch domain.

Table 2.2: The results of 5-way Mixed ANOVA analysis are presented for each factor and the referred interactions.

Factors (or Interactions)	**F-Score**	**Sig.**
Informativeness Measure (Info)	F(7,252)=515.287	p=0.000
Batch Selection Strategy (BS)	F(2.290,105.131)=31.565	p=0.000
Feauture Representation (FR)	F(1,36)=38.964	p=0.000
Seed Set Size (SS)	F(1,36)=8.103	p=0.088
Database (DB)	F(1,36)=150.876	p=0.000
BS * Info	F(21,756)=12.087	p=0.000
SS* Info	F(7,252)=13.060	p=0.000
DB * Info	F(7,252)=180.931	p=0.000
DB * FR * Info	F(7,252)=69.861	p=0.000

F-scores and p-values for 5-way Mixed ANOVA analysis are presented in Table 2.2. Informativeness factor has a significant effect on active learning performance against random learning as shown in Table 2.2. Estimated marginal means for informativeness measures are presented in Figure 2.4. It shows that only the Margin based informativeness measure can confidently perform better than the random selection since its 95% confidence upper bound is lower than the deficiency value of 1. Also note that divergence based methods perform significantly worse than all the other informativeness measures.

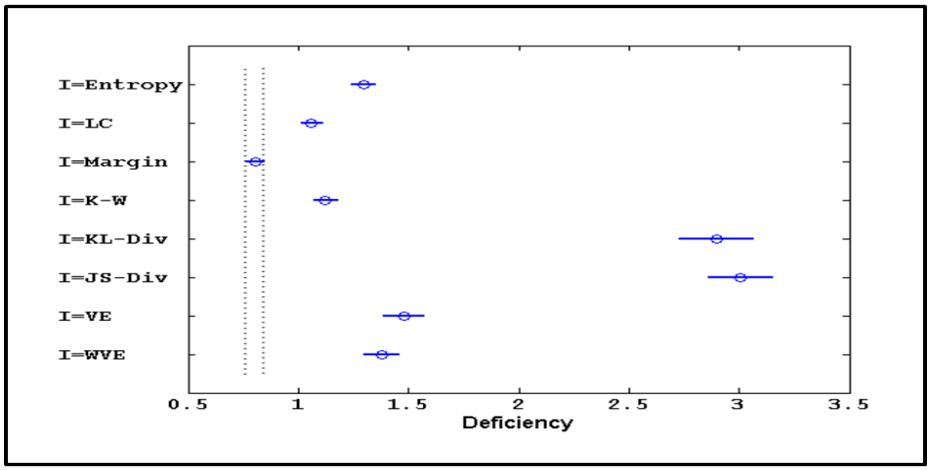


Figure 2.4: The estimated marginal means for informativeness measure factor are presented. The Margin based selection is the only informativeness measure performing significantly better than the random selection.

The Margin based informativeness measure is consistent among databases as shown in Figure 2.5. It confidently performs better than random selection in both databases and there is no significant difference on its performance among databases. Also observe that Körner-Wrobel selection performs as well as the Margin based informativeness measure on the COAD database, but it performs significantly worse than random selection in the Nicicon database. Therefore, the Margin based selection is the only consistent informativeness measure which can perform significantly better than random selection.

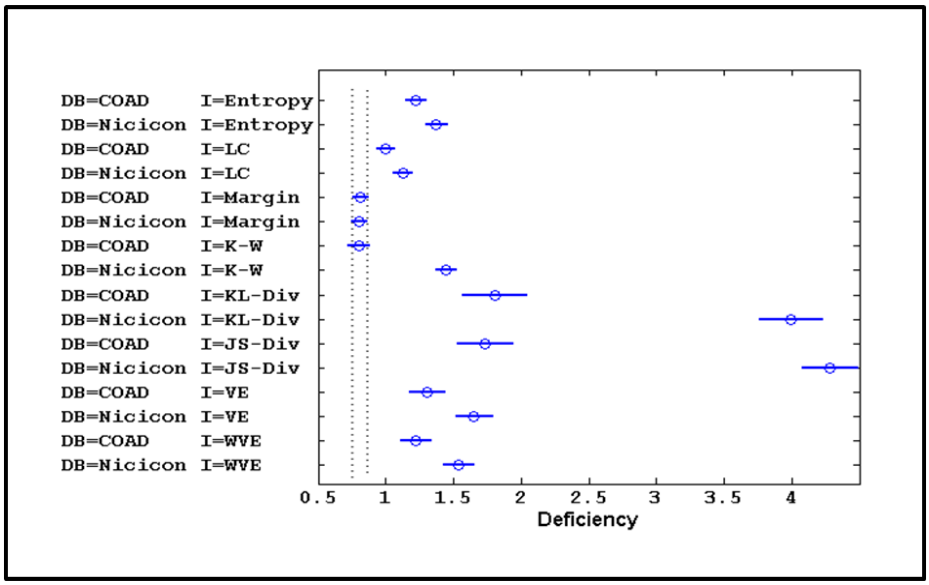


Figure 2.5: The estimated marginal means for the 2-way interaction of informativeness measure and database factors are presented. The significant advantage of the Margin based selection against the random selecion is consistent on both databases.

2.5.2 Choice of Batch Selection Strategy

Due to the re-training cost at each round of the active learning process, users tend to label and add samples as batches rather than adding one sample at a time. This may cause labeling redundant data. Therefore, a proper batch selection strategy is crucial to maintain effective active learning performance. We will present a comparison among basic batch selection strategies along with a detailed investigation on the best performing informativeness measure (the Margin based selection) in this section.

Batch selection strategy (BS) factor significantly affects the active learning performance as shown in Table 2.2. We further investigated performances of the batch selection strategies via Post-Hoc tests and presented the results in Table 2.3.

All the other batch selection strategies perform significantly better than the Default batch selection as shown in Table 2.3. Also observe that Global-FV performs significantly better than all the other methods. Therefore, we can conclude that batch selection strategies increase the performance (over the default performance)

Table 2.3: Bonferroni corrected paired t-test results for active learning methods are presented. Having a mean difference smaller than zero indicates that a method performs significantly better (has a confidently smaller deficiency value) than the reference method. All the other batch selection strategies perform significantly better than the Default selection whereas the Global-FV strategy performs the best among all.

(I) Batch Selection Strategy	(J) Batch Selection Strategy	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Default	Global-FV	.299*	.034	.000	.203	.395
	Combined	.090*	.030	.026	.007	.173
	Global-PE	.122*	.033	.004	.031	.213
Global-FV	Default	-.299*	.034	.000	-.395	-.203
	Combined	-.209*	.030	.000	-.294	-.124
	Global-PE	-.177*	.031	.000	-.264	-.090
Combined	Default	-.090*	.030	.026	-.173	-.007
	Global-FV	.209*	.030	.000	.124	.294
	Global-PE	.032	.031	1.000	-.055	.119
Global-PE	Default	-.122*	.033	.004	-.213	-.031
	Global-FV	.177*	.031	.000	.090	.264
	Combined	-.032	.031	1.000	-.119	.055

and Global-FV is a desirable batch selection strategy.

We further investigate the batch selection strategies on the best performing informativeness measure (the Margin based selection) and present the results in Figure 2.6. Combined strategy performs significantly worse than the other batch selection strategies when used with the Margin based selection. This indicates that the real weight of the informativeness measure is much more than the real weight of the diversity measure and equally weighting of Combined strategy fails with strong informativeness measures. Therefore, the Margin based informativeness measure is compatible with batch selection strategies having high weight on the informativeness measure.

The Margin based informativeness has no significant difference in performance when combined with the Default selection in comparison to the other well performing strategies (Global-FV and Global-PE) as shown in Figure 2.6. Therefore, using Default batch selection with margin based selection yields satisfactory performance and also prevents spending more computation power and time for more complex batch selection strategies.

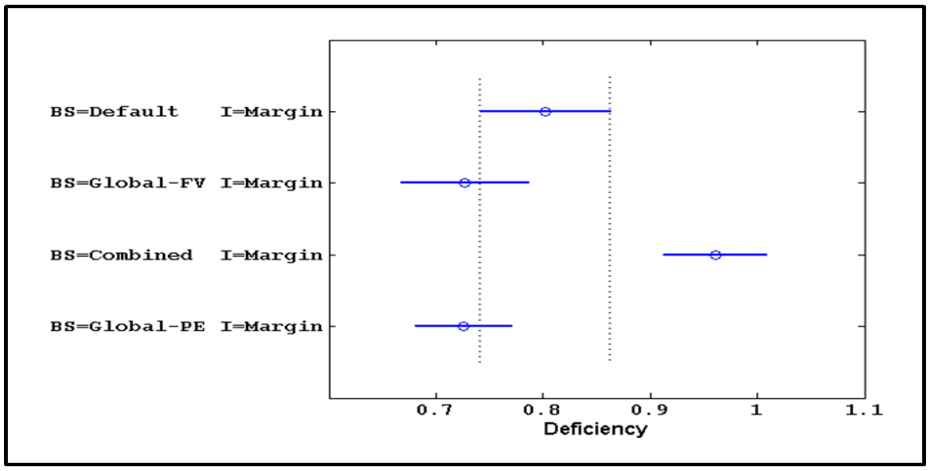


Figure 2.6: The estimated marginal means for the Margin based selection over batch selection strategies are presented. The Margin based selection can have a promising performance even without a sophisticated batch selection strategy.

2.5.3 The Effect of Seed Size

Optimally determining the parameters of the active learning process such as seed size and batch size is an unsolved problem. Although this is the case, our analysis on the effect of seed size yields strong insights to lead the researchers to build working parameter optimization solutions.

Although seed size (SS) factor has no significant effect on the performance of active learning with respect to random selection, 2-way interaction of seed size and informativeness measure factors has a significant effect as shown in Table 2.2. When we further investigate, we observe that initializing with the larger seed size increases the performance of the single classifier based methods with respect to random selection as shown in Figure 2.7. Therefore, a stronger classifier can make more accurate decisions and it prevents selecting misleading samples at the earlier rounds of the process. However the QBC based methods has a little (insignificant) drop in performance. Since we initialize the member classifiers of the committee with the same seed set, they become more alike with a larger seed set and they cannot disagree as much as expected.

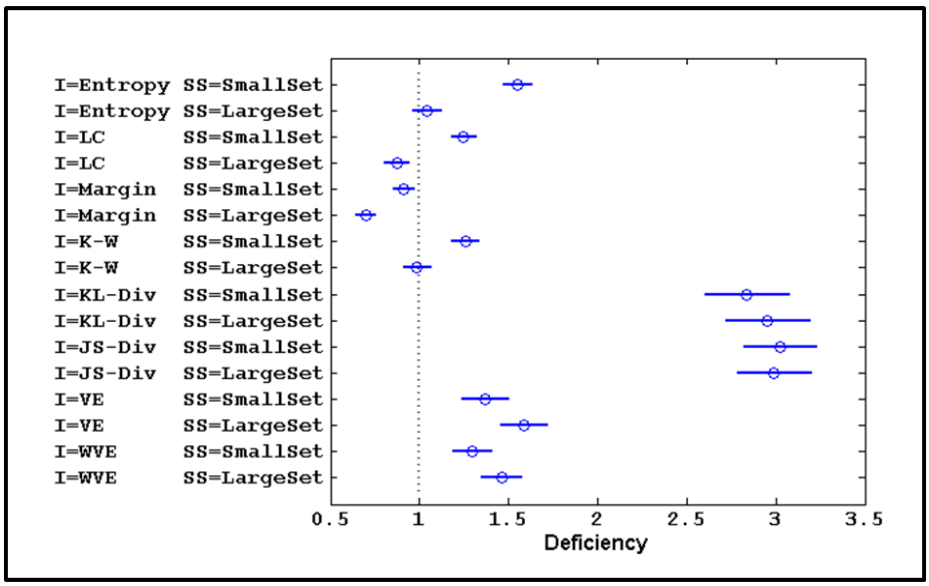


Figure 2.7: The estimated marginal means for the 2-way interaction of informativeness measure and seed set size factors are presented. All single classifier based methods perform significantly better when a larger seed set is utilized.

Table 2.4: For each informativeness measure, estimated marginal mean of the performance gap obtained by using large seed set over small seed set is presented as a result of the 4-way Mixed ANOVA analysis. The Margin based selection is the least sensitive to the choice of seed set size.

Informativeness Measure	**Mean**	**Std. Error**	**95% Confidence Interval**	
			Lower Bound	**Upper Bound**
Entropy	.489	.022	.442	.536
LC	.630	.064	.495	.765
Margin	.846	.098	.639	1.053
K-W	.784	.102	.570	.997
KL-Div	.478	.027	.421	.535
JS-Div	.435	.022	.389	.481
VE	.789	.041	.703	.874
WVE	.785	.055	.668	.901

We also conduct 4-way Mixed ANOVA in order to analyze the performance of active learning for the large seed set against the small seed set. There is confident gain in performance when more seeds are used as shown in Table 2.4. In addition, the Margin based informativeness has a deficiency value not confidently less than 1 which indicates that using the large seed size does not have a significant performance gain with respect to using the small seed size. Therefore, the Margin based informativeness is the least sensitive measure to choice of seed size. Also Global-FV and Global-PE increases the resistance to improper choice of seed size as shown in Table 2.5. Therefore, batch selection strategies can be utilized to reduce the effect of the choice of seed size.

Table 2.5: Bonferroni corrected paired t-test results on the estimated marginal means of 4-way Mixed ANOVA analysis. Having a mean difference smaller than zero indicates that a method has significantly more resistance to the degrading effects of seed size choice (has a confidently smaller deficiency value on the performance gap) than the reference method. Employing Global-FV or Global-PE strategies reduces the effects of the choice of improper seed set size significantly more than the Default selection

(I) Batch Selection Strategy	(J) Batch Selection Strategy	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Default	Global-FV	-.158*	.042	.009	-.282	-.033
	Combined	-.008	.027	1.000	-.087	.071
	Global-PE	-.124*	.029	.003	-.209	-.038
Global-FV	Default	.158*	.042	.009	.033	.282
	Combined	.150*	.024	.000	.080	.220
	Global-PE	.034	.022	.844	-.031	.099
Combined	Default	.008	.027	1.000	-.071	.087
	Global-FV	-.150*	.024	.000	-.220	-.080
	Global-PE	-.116*	.015	.000	-.159	-.072
Global-PE	Default	.124*	.029	.003	.038	.209
	Global-FV	-.034	.022	.844	-.099	.031
	Combined	.116*	.015	.000	.072	.159

2.5.4 Utilizing Prior Knowledge

Having prior knowledge on the database and the feature representation can give hints when utilizing active learning. Since the Margin based selection is the only measure

outperforming random selection, we investigate only its performance under two-way interaction of database and feature representation factors.

The most important deduction is not to use active learning if the feature representation is weak or untrustable (e.g. a new feature set which is unjustified). Note that for a feature representation, being weak is with respect to the database. Although the performance of the Margin based selection is not significantly affected by the feature representation on the COAD database, its performance on the Nicicon database is significantly low with Zernike features with respect to its performance with IDM features. Figure 2.8 depicts this observation.

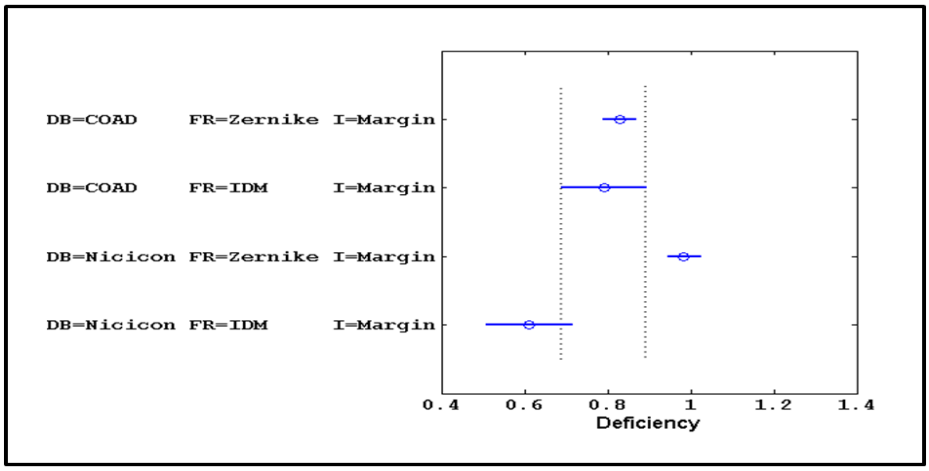


Figure 2.8: The estimated marginal means for the Margin based selection over 2-way interaction of database and feature representation factors are presented. This implies that we should not rely on active learning when the feature representation is weak and the database is "hard" to learn.

The other important deduction is that being an easier database suggests that the samples are representative and informative by themselves under a strong feature representation. Therefore, random selection can work as well as active learning on these databases. When IDM features are used, the active learning performance is almost significantly higher on the Nicicon database than the COAD database. Since the COAD database has less noise and style variation than the Nicicon database has, its samples are more representative.

In conclusion, when the database is too simple and the feature representation is too strong; or when the database is too hard to learn and the feature representation is too weak, active learning might fail. Thus, relying on active learning on such cases might be more costly than using random learning.

2.6 Discussion

We expect our analysis to serve as an example for researchers who might want to apply active learning in sketch recognition domain, or in other domains. Our experiments were on sketch domain due to the databases and feature representations used. Although this is the case, the insights obtained from the analysis are applicable to any active learning process.

We investigated a wide-range of informativeness measures and batch selection strategies. We showed the consistently best performing informativeness measure is the Margin based informativeness mesaure for sketch recognition. It is also desirable for not requiring a batch selection strategy since employing a batch selection strategy did not yield a significant performance difference for the Margin based selection. Therefore, it can be confidently used without a batch selection strategy when the batch size is not large. Besides, it is the most insensitive informativeness measure to the choice of seed set size so that it saves the users form worrying about the sufficiency of the seed set with respect to the other measures.

We showed that batch selection strategies increase the active learning performance significantly when we observe the performances of the informativeness measures all together. We observed that combining an informativeness score with another score (such as diversity score) through weighing the scores by a constant might be risky for working (powerful) informativeness measures (such as the Margin based selection). Unless we have a notion of how to weigh informativeness score with another score, we should stay away from such batch selection strategies. Applying batch selection strategies over unlabeled samples with top informativeness scores might be safer as our results (on Global-FV and Global-PE strategies) indicates. However, we need to

determine which unlabeled samples are the ones with the top informativeness scores. We employed a small constant based on the number of classes of employed database, in our experiments. However, this process should be more dynamic such that a cut on the sorted set of informativeness scores is searched. Another option will be applying additive increase and multiplicative decrease strategy on the limit of the number of unlabeled samples with top informativeness scores considered. In such a strategy, we might start with a low value for this limit and decide to increase the batch depending on the observations such as the validation accuracy at each round and properties of clusters of informativeness scores when sorted.

One crucial side of our analysis is that we investigate active learning methods over several factors. Hence, we are able to report results on the factors effecting the active learning performance. One important factor in the performance of active learning is the seed set size. We showed that initializing the process with a larger seed set yields better active learning performance for single classifier based methods. This is not valid for QBC based methods since we initialize the committee members with the same seed set. Hence, as the committee members gets similar with more (shared) seeds, their decisions made by disagreement measures get less reliable. Therefore, design decisions for QBC based methods should consider initialization procedure carefully. We have the insight that initial classifier should not be too weak because predicted class probabilities for unlabeled samples become garbage and detailed methods (such as entropy and divergence based methods) further increase this garbage. Hence, utilizing a detailed informativeness measure with too weak initial classifier causes selection of ineffective samples and yields very poor performance for earlier rounds of active learning process which are the crucial rounds to overperform random selection.

We suggest active learning users to utilize the prior knowledge on their database and feature representation. Simply, if they do not trust their feature representation or if they use a new feature representation, it is risky to employ active learning. Our results indicate that if the database is "hard" to learn, active learning performance is poor with a weak feature representation. Also if samples in a database are too much

representative, using active learning is unnecessary. If resources allow, we suggest to apply a few trial rounds both with active learning and random learning in order to make a more accurate decision on whether to utilize active learning or not when we do not have prior knowledge on the database and feature representation.

We encourage future researchers to employ factor based analysis in their studies to have a deeper understanding of active learning methods and the process of active learning.

Chapter 3

ACTIVE SCENE LEARNING

3.1 Introduction

We can categorize active learning into two frameworks depending on the properties of queried instances: "instance based" active learning and "multi-instance based" active learning. The main difference is that multi-instance based active learning targets structured instances or bags of instances rather than seeing instances as a single entity. For example, a sentence in a sequence labeling task is a structured instance (of tokens). In a similar way, in a content-based image retrieval task, an image is a bag of instances such that the image is over-segmented and these segments all together constitute a bag.

There are various active learning methods proposed on instance based and multi-instance based active learning frameworks. Although these methods can be employed in many tasks (e.g. image segmentation, sequence labeling, image categorization and text categorization), we cannot apply these methods directly to sketch recognition due to some properties of sketch data. For sketch recognition, we construct fragments from ink data (with no color or intensity information) and then, our primitives are fragments. However, they do not have enough information themselves to be labeled individually. We can give them a meaning only when they are considered in a group (of fragments). Moreover, definition of segmentation separates our task from other segmentation tasks (e.g. image segmentation).

Segmenting an instance by assigning labels to its primitives (e.g. pixels or super-pixels) is implicit segmentation, since the segmentation will be automatically obtained after labeling the primitives is completed. Therefore, such idea of segmentation aims to determine to which specific object class a primitive belongs. However, in our task,

grouping the collection of primitives into separate objects is the process of segmentation; and object-level labeling of these groups of primitives is the process of labeling. This definition of segmentation is valid for any sequential data. For example, we might want to segment a video into activities/gestures. If we divide the video into windows, labeling each window without considering it in a group will be extereme to achieve a correct segmentation. In such a case, a dynamic programming solution considering groups of these windows will yield an optimal solution. Definition of segmentation and instance properties is crucial for working mechanism of active learning methods. Therefore, for sequential data and for sketch recognition (strictly speaking for non-interspersed sketches), we need a new active learning framework.

We will now present an example usage of active learning for some tasks and provide a better intiution of why we need a new active learning framework. For example, in image segmentation, informative pixels (or super pixels) are detected and labeled as regions or clusters [21, 22, 23, 24, 25]. This is mostly an iterative process where the user is requested to label new regions of pixels on a single image. However, we want a framework to improve a specific model for segmentation and recognition rather than segmenting a specific sample.

Similar to image segmentation, sequence labeling has implicit segmentation [7, 26]. Tokens are labeled in a sentence where a sentence is an instance. In this task, the active learning method selects samples by taking token-wise or sentence-wise confidences into account. However, sketches do not have pre-defined meaningful parts, but fragments. Therefore, we need to generate and consider possibly meaningful parts which are the candidate segments obtained by segmenting the sketch with the existing (current) model, during the active learning process. Besides, a proper framework should allow providing information for correct segmentation as well as labels.

Another approach is the "bags approach" which is used for image and text categorization [27, 2, 28, 29, 30]. In this approach, an instance can be an image of an object class or a segment belonging to an object class in the multi-instance active learning framework. This approach considers whether there is an instance of a specific object

class in the bag or not. Therefore, the bag approach is limited with a binary class labeling scenario compared to multi-class availability for which we seek. We also use the term "scenes" rather than bags in this aspect.

We propose a framework which distinguishes from the usual instance based and multi-instance based frameworks in terms of the segmentation aspect, properties of the training instances and multi-class labeling capability for the scenes. Our framework also supports labeling at mixed granularities such that a user can label a whole scene or a part of the scene. In addition, we adapted several multi-instance based methods to our framework and analyzed their performances. Moreover, we proposed the method of segment-wise active learning on scenes and showed its strength via an empirical analysis.

This chapter is organized as follows: First, we introduce our framework. Then, we propose scene-wise and segment-wise selection schemes in Section 3. We explain the experimental design and analysis methodology in Section 4 and Section 5. We demonstrate the analysis results in Section 6 and conclude the chapter with a discussion.

3.2 Active Scene Segmentation Framework

3.2.1 Description of the Framework

The aim of our framework is to carry out active learning to incrementally improve models targeting segmentation and labeling of samples consisting of objects from various object classes. Our framework enables active learning on systems such that grouping primitives into separate objects in a collection of primitives is the process of segmentation and object-level labeling of the groups of primitives is the process of labeling. Therefore, in our framework, active learning methods should request segmentation and labels of objects explicitly rather than requesting labels to denote to which object class a primitive belongs. This is counter-intuitive for generative model based systems such that each primitive has a meaning (or label) on its own.

Our framework allows querying instances consisting of objects from various object

classes unlike the "bags approach". Moreover, we can do labeling at mixed granularities since it is possible to create active learning methods querying on both scene-level and segment-level in our framework.

We initialize an active learning process by training the model/system with a few labeled data, the so called "seed set". Seed set contains instances of isolated objects from each object class. The process continues in rounds until available resources are consumed or a target validation accuracy is reached. A round starts by segmenting and classifying unlabeled scene instances in the pool. At this point, our framework allows various schemes to measure informativeness of scenes or partial scenes. We categorize these schemes as scene-wise and segment-wise methods. We describe them in detail, in the following sections. Depending on which scheme is employed, user corrects segmentation and provide labels either for all the objects in the scene or for only the targeted objects. Then, the round ends with adding newly labeled objects to the training set and re-training the model.

3.3 Active Scene Learning

We approach scene labeling in two ways: scene-wise and segment-wise. Scene-wise selection scheme targets to compute an overall informativeness score for scenes, according to individual informativeness scores of the candidate segments in the scenes. In addition, it requests users to provide corrections for all mis-segmentations and mislabels in a selected scene. Unlike scene-wise selection scheme, segment-wise selection scheme is "targeted" so that it only requests corrections for the objects intersecting with the most informative candidate segment(s).

3.3.1 Scene-wise Active Learning

After segmenting and recognizing unlabeled scenes in the pool, each segment belonging to the most likely scene segmentation is called as a "candidate segment". In order to compute individual informativeness scores for candidate segments, we employed Margin based informativeness measure (as introduced in Chapter 2). In order to

assess an overall informativeness score for a scene, we applied basic operations over the individual informativeness scores of candidate segments in a scene. These basic operations include "arithmetic mean", "geometric mean" and "max". We also use estimated scene probability (probability of the most likely segmentation predicted) in addition to these basic methods. These methods are also applied in multi-instance based active learning and on several domains [7, 2, 28, 30, 31, 32]. Hence, we adapted these basic methods to our framework and evaluated on our domain. Now, we will present the scene-wise selection methods:

Arithmetic Mean Approach: Roth and Small proposed several active learning methods on structured output spaces [32]. We adapted their Q_{local} method on our framework. After segmenting and recognizing an unlabeled scene, we compute Margin based informativeness score for each candidate segment in the scene. Then, we compute the mean of these informativeness scores and assign the resulting score to the scene. Then, we sort the scenes depending on this final score. Finally, we choose a batch of scenes with the highest scores.

Geometric Mean Approach: This method follows the same procedure as in the arithmetic mean approach. However taking a geometric mean penalizes scenes more than taking an arithmetic mean. Even if the scene has only one uninformative segment (which has informativeness score close to zero), it drastically affects the overall score of the scene.

Maximum of the Scene Approach: This method is similar to Roth and Small's $Q_{local(C)}$ method [32]. We compute margin based informativeness score for each candidate segment in a scene. Then, we assign the score of the most informative candidate segment to the scene. Then, scenes are sorted depending on this score and a batch of scenes with the highest scores are chosen.

Scene Probability Approach: One minus the probability of the most probable interpretation of the scene is assigned as the informativeness score of the scene. This method is similar to the least confident based selection proposed by Settles and Craven on the sequence labeling task [7]. They compute the probability of the most likely

sequence labeling and subtract this probability from 1 to achieve informativeness of the sequence.

Algorithm 1 Algorithm for Scene-wise selection

```
Initialize the model with the seed set.
repeat
    Segment and recognize the scenes in the pool with the current model.
    Compute scene informativeness scores with a desired scene-wise selection method.
    repeat
        Select the most informative unlabeled scene.
        Request user to provide segmentation and labeling for all the objects in the scene.
        Add all the objects in the scene to the training set.
        Remove the scene from the pool.
        Assign the scene with the next highest informativeness score as the most informative scene.
    until the batch is full
    Re-train the model with the enlarged training set.
until the halting point is reached
return the model
```

3.3.2 Segment-wise Active Learning

Scene-wise active learning methods force users to provide full segmentation and full labeling to the selected scenes. A selected scene might contain uninformative objects, especially for the later rounds. The effort for providing segmentation and labeling for these objects is redundant. Therefore, we propose a simple and efficient method to overcome this problem.

We obtain candidate segmentation for all unlabeled scenes in the pool by segmenting and recognizing these scenes with the current model. Then, we compute the

Margin based informativeness scores for each candidate segment for each scene in the pool and sort these individual scores of the candidate segments. By starting from the scene having the highest scored segment, we add each object that the candidate segment touches on. We depicted the labeling for a scene with the most informative candidate segment via an example in Figure 3.1. There are k objects and N fragments in the scene. We provide both an actual segmentation and a candidate segmentation for the scene, for demonstration purpose. First two objects in our example are mis-segmented. In addition, the candidate segment enclosed with a blue rectangle represents the most informative candidate segment in the pool. Therefore, user is requested to label this scene. However, we ask user to label only $Object_1$ and $Object_2$ rather than all the objects in the scene. Therefore, user provides segmentation and label only for informative objects.

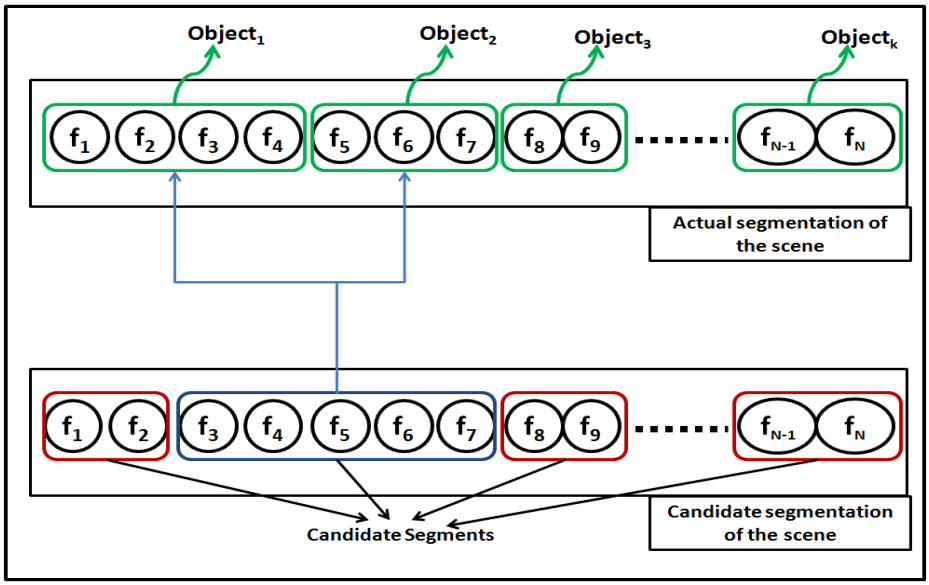


Figure 3.1: Visualization of Segment-wise selection scheme.

We also present an actual example with a real data. Figure 3.2 represents a scene composed of 3 objects from the Nicicon database. We colored the figure to represent the candidate segmentation given by the current model. Each color refers to a candidate segment. First two objects are mis-segmented whereas the last(most

right) object is correctly segmented. Red colored candidate segment contains the head of the stick figure which is the second object. Assume that red colored candidate segment has the highest informativeness score in the pool and user is asked to label this scene. User should correct mis-segmentation and mis-labeling for the "causality" symbol (leftmost object) and the "injury" symbol (the lying stick figure). Then, the scene containing the candidate segment with the next most highest informativeness score is presented. This process continues untill batch gets full.

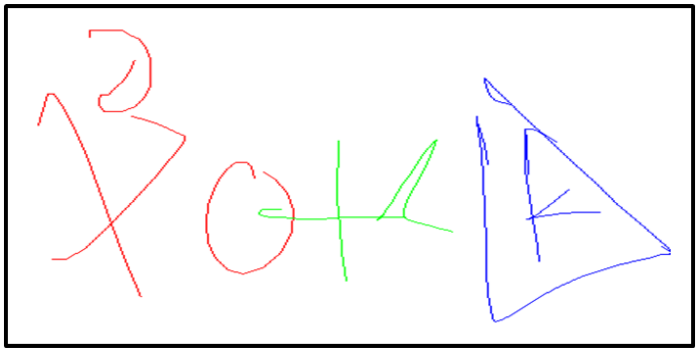


Figure 3.2: A sample visualization of a candidate segmentation for a scene with 3 objects from the Nicicon database.

We add the feature vectors of the labeled objects to the training set. However we do not modify the scene after labeling. Therefore, a scene stays the same after being partially or completely labeled. Although it is unlikely for a labeled object to be informative again, it can appear in an informative candidate segment. However we do not add the same objects to the training set throughout the active learning process. We have a simple check for this and we pass to the next highest informative object, in such a case.

A batch of objects is added to the training set in each round. If the number of the objects exceeds the batch size after correcting a candidate segment, the labeling process ends and re-training is done.

Algorithm 2 Algorithm for Segment-wise selection

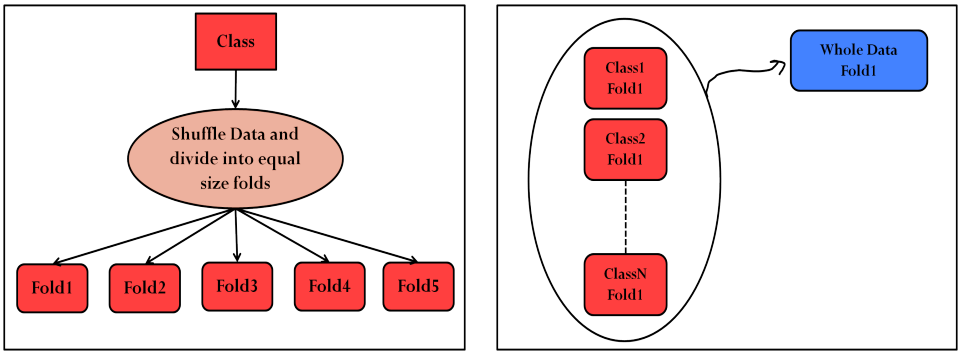
```
Initialize the model with the seed set.
repeat
    Segment and recognize the scenes in the pool with the current model.
    Compute the Margin based informativeness score for each candidate segment.
    repeat
        Select the scene containing the candidate segment with the highest informativeness score.
        Request user to correct mis-segmentation and mis-labeling for the objects intersecting with this candidate segment.
        Add the labeled objects to the training set if they are not already added.
        Assign the candidate segment with the next highest informativeness score as the most informative candidate segment.
    until the batch is full
    Re-train the model with the enlarged training set.
until the halting point is reached
return the model
```

3.4 Experimental Design

3.4.1 Databases and Feature Representations

We created synthetic scene datasets from available individual symbol sketches. In this section, we first present source databases from which we created scene datasets, then we describe how we create synthetic scene datasets.

We have two source databases. The first database includes symbols from our Course of Action Diagrams (COAD) database [16], and the other database is the publicly available Nicicon database [17]. The COAD database contains 620 samples and 20 different classes whereas the Nicicon database contains 22958 samples and 14 different classes.

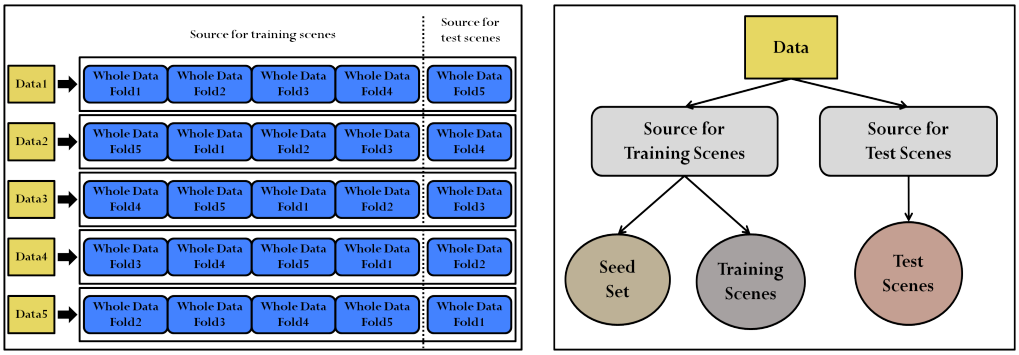


(a) (b)

Figure 3.3: The process of source data creation for synthetic scenes. (a) Data of each class is divided into folds. (b) Merging the corresponding folds of each class into larger folds.

We created scene datasets by considering proper statistical analysis. For each source database, we applied the following procedure.

- Create 5 equally sized folds for each class in a random manner (see Figure 3.3a).
- Merge respective folds of each class and obtain larger folds with objects from each class (see Figure 3.3b).



(a) (b)

Figure 3.4: The process of (a) creating source data for 5 repeats and (b) utilization of source data.

- Create source data for training and test scenes for 5 repeats (see Figure 3.4a). For each repeat:
  - A different fold is separated as source for test scenes.
  - The rest of the folds are merged to obtain source for training scenes.
- Select 4 samples from each class in a random manner to obtain seeds. Then, remove the selected samples from the source for training scenes (see Figure 3.4b).
- To create a scene with k objects, apply the procedure depicted in Figure 3.5.

We created various sized scenes which contain 2, 3, 4, 5 and 6 objects. There are 20 instances for each size option for both training and test sets. Therefore, for a repeat, we ran the tests with 100 training scenes and 100 test scenes along with 4 seeds from each object class. We have 5 repeats of a trial for each source database.

We employ IDM features [19]. Parameters of IDM feature extraction are kernel size, sigma and resample interval (with values 25, 2 and 100 respectively). The length of the feature vector is 720 with the given parameters for IDM.

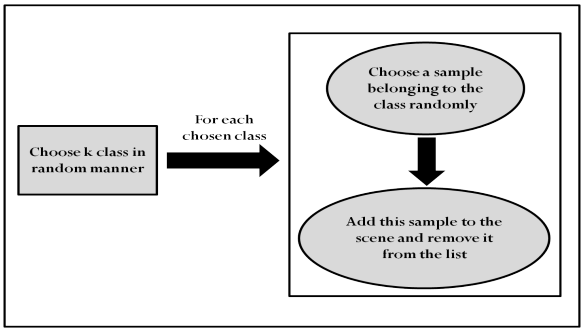


Figure 3.5: The process of generating a single scene with k objects.

3.4.2 Trials

We experimented with 4 scene-wise selection methods (GM, ArM, SP, MoS), segment-wise selection method and random selection on scene datasets created from 2 different source database. We repeated the trials 5 times with synthetic datasets which has no overlap on their test data. For each repeat, we sustained active learning process 25 rounds for each trial. For each round, we added 3 scenes (12 objects in average) for adapted methods and 12 objects for our method.

We have a pre-process step before initialization. The offline pre-process saves time and computation during classification of the pool elements and acquiring test results at each round. We fragmented both train and test scenes before starting the active learning process. In addition, we extracted features for the possible combinations of fragments in the scenes which will be used to fill the DP matrix during segmentation of the scenes at each round. These possible computations are determined by Imin and Imax parameters of the segmentation and recognition module. These parameters indicate how many fragments an object can have at least (I_{min}) and at most (I_{max}). In our evaluation, I_{min} is 1 for both the COAD and the Nicicon databases whereas I_{max} is 20 for the COAD database and 24 for the Nicicon database.

3.4.3 Description of Segmentation and Recognition Module

We work on sketch domain. Sketch data does not have color or intensity information, but ink data. We keep position and time information for each ink dropped on the canvas. Our sketch recognition module aims to solve non-interspersed sketches such that the user must finish drawing an object before starting to draw another object.

In order to segment and recognize a scene, we first fragment the scene. Therefore, our primitives become the fragments and we can concentrate on grouping the fragments into objects and predicting object-level labels. We fragment the scene with Tumen's method [33]. Then, we do segmentation and recognition by applying Sezgin's method [34].

From the fragments, Sezgin constructs a graph $G(V, E)$ in which vertices V correspond to the fragments indexed by the order in which they were drawn. The weight $w(i, j)$ associated with an edge from v_i to v_j in G corresponds to the probability that the set of fragments between i and j corresponds to a valid and fully-drawn symbol where $P_{i,j}(C)$ corresponds to the probability that the set of fragments between indices i and j corresponds to class C. The optimal segmentation is computed through dynamic programming where $S(i, j)$ is the optimal segmentation of a sub-sketch made of fragments from indices i to j:

$$S(i,j) = \max \begin{cases} \text{w(i,j)} \\ \max_{i \le k < j}(S(i,k).S(k,j)) \end{cases}$$

Sezgin employs a multi-class SVM to compute $P_{i,j}(C)$ values. This classifier is supported by a one-class SVM. If a set of fragment is classified as "Non-object" by the one-class SVM, $P_{i,j}(C)$ value is set to zero for that set of fragments. However this penalty is high and propagates through dynamic programming. Although we employed grid search including ν parameter, and also tried the best values of ν parameter in Sezgin's article [34], accuracy of one-class SVM was not sufficient enough to yield a satisfactory segmentation performance in the end. Hence, using multi-class SVM alone is more accurate and robust. Therefore, we use only the Multi-Class SVM in our experiments.

3.5 Analysis Methodology

We use the deficiency metric described in Section 2.4.1. Performance curves used for computing deficiency values have number of labeled (isolated) objects on their x-axis and accuracy values obtained in each round on their y-axis, similar to isolated sketch recognition case. However, we have the segmentation and recognition accuracies on the y-axis for scenes. It is the ratio of the number of correctly segmented and labeled (individual) objects over the total number of individual objects (over all test scenes). Note that in order to correctly label an object, it must be segmented correctly.

In order to assess the statistical significance of the differences observed in the deficiencies obtained from different active learning methods, we conducted 2-way Mixed ANOVA analysis. Throughout our analysis, we performed Mauchy's sphericity test to check whether the variances of the differences between all possible group pairs, subject to ANOVA, are equal. If sphericity is violated, the degrees of freedom are corrected using Greenhouse-Geisser correction. We also performed Levene's test for checking homogeneity of variances between groups and used transformed values where appropriate. Bonferroni corrected paired t-tests were performed as Post-Hoc tests.

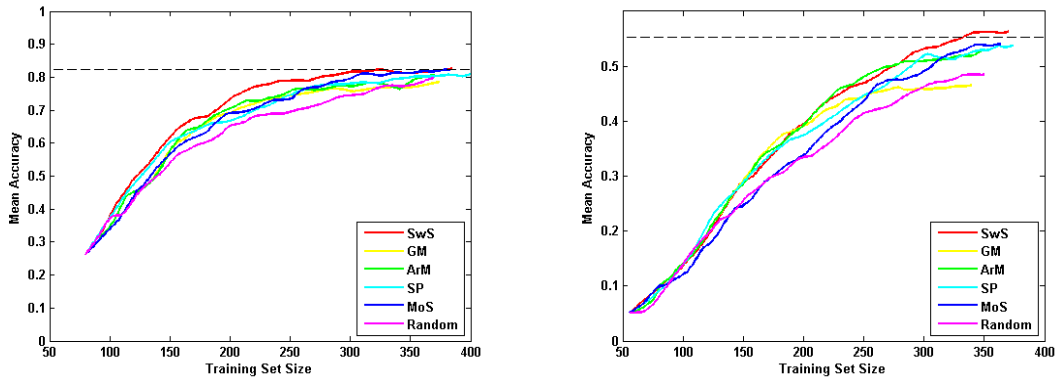
The database factor is the between variable factor and the active learning choice is the within variable factor in our 2-way mixed ANOVA design. The dependent variable is the deficiency values of the active learning methods computed with respect to random selection results. Therefore, the performance of the random selection stands as a reference point for comparing the performances of active learning methods. In addition, the deficiency metric has implicit comparison of the performances of any active learning method and the random selection.

3.6 Results

The (labeling effort) gain with respect to the performance of random selection is crucial, in the concept of active learning. Before presenting our promising results on this aspect, we will also point out remarkable gains of active learning against training

with the whole data.

We present mean accuracy graphs of active learning methods and random selection for two source databases in Figure 3.6. The dotted line represents the mean accuracy obtained when we train the model with the whole training data (and seeds). The synthetic datasets created from the COAD database contain 80 seeds and 400 unlabeled individual objects over 100 scenes in total. After initializing with 80 seeds, our segment-wise selection scheme (SwS method) selects 243 samples and reaches the accuracy of training with 480 samples (as shown in Figure 3.6a). This is a % 40 gain over 400 unlabeled samples. The synthetic datasets created from the Nicicon database contain 56 seeds and 400 unlabeled individual objects over 100 scenes in total. After initializing with 56 seeds, our SwS method selects 281 samples and reaches the accuracy of training with 456 samples (as shown in Figure 3.6b). This is a % 30 gain over 400 unlabeled samples. Therefore, our SwS method avoids ineffective data which has no influence on the model.



(a) (b)

Figure 3.6: Mean (segmentation and recognition) accuracy graphs for the synthetic scene datasets generated from (a) the COAD database and (b) the Nicicon database. Active learning methods outperforms random selection. Also our SwS method outperforms adapted multi-instance based methods, especially on the COAD database.

We can observe that random selection performs poorly with respect to active learners as seen in Figure3.6. We will verify the significance of this observation via the

results of ANOVA analysis. In addition, we demonstrate the performance comparison among active learning methods via Post-Hoc tests of ANOVA analysis.

Table 3.1: Estimated marginal means for active learning methods on deficiency metric. SwS, ArM, SP and MoS methods perform confidently better than random selection.

AL_method	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
SwS	.747	.033	.670	.824
GM	.894	.051	.777	1.011
ArM	.833	.049	.720	.946
SP	.846	.029	.779	.913
MoS	.896	.032	.822	.971

All active learning methods, except GM, perform confidently better than random selection as shown in Table 1. For all active learning methods, except GM, upper bound of the 95% confident interval is less than 1. This indicates that each of them has a deficiency value confidently less than 1. Therefore, they confidently perform better than random selection. Also our segment-wise selection method is the most confident among all. Observe that GM and ArM have high variance and this is one of the reasons that they have a higher upper bound value. We might imply that their performance may not be stable.

Table 3.2: F-scores and p-values for the factors of 2-way Mixed ANOVA analysis. The choice of active learning method has a significant effect on the performance whereas the choice of database has no significant effect on the performance.

Factor	F-Score	Sig.
AL_method	F(2.136,17.085)=10.614	p=0.001
Database	F(1,8)=0.619	p=0.454

We present F-Scores and p-values of the factors of our (2-way Mixed) ANOVA analysis in Table 2. AL_Method factor has a significant effect on the performance.

Thus, choice of active learning method matters. The database factor does not have a significant effect. This indicates that performances of active learning methods are consistent among databases.

Table 3.3: Bonferroni corrected paired t-test results for active learning methods. Having a mean difference smaller than zero indicates that a method performs significantly better (has a confidently smaller deficiency value) than the reference method. Our SwS method outperforms GM, SP and MoS methods and performs almost significantly better than ArM method.

(I) AL_method	(J) AL_method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
SwS	GM	-.147*	.035	.032	-.283	-.011
	ArM	-.086	.034	.356	-.218	.045
	SP	-.099*	.021	.014	-.179	-.020
	MoS	-.150*	.023	.002	-.237	-.062
GM	SwS	.147*	.035	.032	.011	.283
	ArM	.061*	.015	.042	.002	.120
	SP	.048	.026	1.000	-.053	.148
	MoS	-.003	.027	1.000	-.106	.101
ArM	SwS	.086	.034	.356	-.045	.218
	GM	-.061*	.015	.042	-.120	-.002
	SP	-.013	.024	1.000	-.104	.078
	MoS	-.063	.031	.775	-.184	.057
SP	SwS	.099*	.021	.014	.020	.179
	GM	-.048	.026	1.000	-.148	.053
	ArM	.013	.024	1.000	-.078	.104
	MoS	-.050	.020	.379	-.128	.027
MoS	SwS	.150*	.023	.002	.062	.237
	GM	.003	.027	1.000	-.101	.106
	ArM	.063	.031	.775	-.057	.184
	SP	.050	.020	.379	-.027	.128

We compare active learning methods according to their performances with respect to random learning. This meaning is implicitly brought by the deficiency metric. To further observe the ANOVA results, we conducted Bonferroni corrected paired t-tests for active learning methods and present the results in Table 3. Our SwS method has a significant mean difference against GM, SP and MoS methods. This indicates that our SwS method has a deficiency value confidently lower than GM, SP and MoS methods as highlighted in Table 3. Therefore, users can select our SwS method over GM, SP and MoS methods and acquire better performance. In addition, our SwS method performs better than ArM method as well, but significance is not observed due to the high variance of ArM method.

3.7 Discussion

We presented a new active learning framework to address tasks including segmentation and multi-objects from various classes per instance. Also our framework requests object-level labeling rather than primitive-level labeling. Therefore, our framework targets explicit segmentation rather than implicit segmentation. By explicit segmentation we mean that primitives are grouped into objects and then they got labeled whereas implicit segmentation means that labeling all primitives in an instance yields the segmentation result itself. Therefore, our framework is required for data composed of primitives which have no meaning by themselves, but have a meaning only considered in a group of primitives.

We introduced two selection schemes: scene-wise and segment-wise. Scene-wise selection methods are adaptation of the basic methods used in previous frameworks whereas segment-wise selection is more specific to our framework. We showed that each active learning method we analyzed under our framework performed significantly better than the random selection (except GM method). We also compared all active learning methods among each other and our segment-wise selection strategy (SwS method) performed significantly better than all scene-wise selection methods, except ArM method (no significance observed due to high variance of ArM method).

Throughout our analysis, we observed that taking a geometric mean to compute an overall informativeness score from individual informativeness scores induces more penalty for a scene rather than taking an arithmetic mean when one or more individual informativeness score is low in the scene. We expected this to allow us select very informative scenes since only the scenes with very informative candidate segments would be chosen. However, even if one candidate segment has a confident recognition, informativeness score of the scene gets closer to zero with geometric mean approach. Hence, the number of the scenes with informativeness scores closer to zero increases as the model gets more powerful after each round. Therefore, we lose information while we are seeking for more information. As it can be seen from the mean graphs, geometric mean approach cannot select samples affective to increase the accuracy

after half-way of the process is reached (this observation is more visible for Nicicon database).

Rather than computing overall informativeness scores for scenes and tolerating uninformative candidate segments, we target to select the full-objects intersecting with the most informative candidate segments. Note that this approach is not the same as the active learning approach using point-wise informativeness scores while segmenting an image since we request labels for objects, not for the primitives. Also we request user to provide correction for segmentation. In addition, this correction is not only directed inside the so called informative area (candidate segment), but around of this area. Hence, a candidate segment also encodes information from the mistakes done around of it.

As a final note, we experimented on non-interspersed sketches. However, our framework and selection approaches can be applied to any task where segmentation module returns a list of objects with object-level labels rather than a list of primitives with labels indicating from which object class they are.

Chapter 4

RELATED WORK

Active learning is a newly established area in the machine learning community. Main reason for such a field to emerge is rapidly increasing amount of available data and information extraction methods founded in parallel. Since both personal storage devices and web repositories have reached extensive storing capacities, people can create and share enormous amounts of data over the web. The increase in the accesibility of the internet and the social networks allow and attract people to share more data. Availability of data paves the way for many applications employing machine learning. Therefore, a need for reducing labeling effort occurs in order to utilize this excessive amount of data. Hence, active learning takes to the stage.

Although active learning field has theoretically well-founded methods, they are not sufficient alone in practice. Therefore, researchers should focus more on the factors affecting the active learning performance. In this aspect, we conduct an extensive empirical analysis for invesitigating such factors. As the related work, we will compare our empirical analysis with the existing analyses from various aspects in the following section.

We propose a new active learning framework and metods under this framework in this thesis. Although we provided differences of our framework from the existing frameworks in Chapter 3, we will now provide a more detailed comparison with the existing frameworks. We will direct our attention on the active learning methods aiming segmentation and the methods employing multi-instance based approach.

4.1 Empirical Analyses in Active Learning

To our knowledge, our work is the first extensive empirical analysis of active learning methods in the sketch domain. We present detailed analysis on factors such as informativeness measure, batch selection strategy, feature representation, seed size and database. We also provide how prior knowledge on the domain can be utilized for proper use of active learning in addition to our extensive comparison of combinations of informativeness measures and batch selection strategies.

Many basic and task specific active learning methods have been proposed in the literature. Olsson presents basic informativeness measures and approaches to active learning in a literature survey [9]. This survey also contains various concerns on active learning such as data access, re-use of annotated data, cost sensitive design and monitoring performance. In another literature survey, Settles presents query strategy frameworks and practical considerations including batch selection, noisy oracles and variable labeling costs [6]. Although both literature surveys are detailed and well designed, there is no empirical study aiming to analyze these methods extensively. In that respect, our study closes this gap in the active learning literature.

Most of the empirical analyses in the active learning literature intend to compare the performance of the newly proposed method with a few baseline methods [4, 5, 7, 11, 14, 21, 22, 23, 25, 30, 32]. Although these analyses are valid, they will be more complete if they provide a factor analysis and a discussion on the effects of the factors. Such discussions will be guiding for future researchers on the area. As we know more on the factors affecting the performance, we will be more aware while creating new active learning methods.

Settles and Craven analyze various basic methods with their adaptations for sequence labeling task all together [7]. They have information density, expected gradient length, and Fisher information in their performance analysis as well as some of the basic strategies. Schein and Ungar analyze the classifier certainty method, as well as several methods of the single classifier approach and the QBC approach for logistic regression [3]. Rather than analyzing various basic methods in the litera-

ture, Markowitz analyzes variations of uncertainty sampling for large corpus labeling with boosted naive Bayesian style classifier [35]. Upon previous empirical analyses, our work presents an analysis of more basic methods and their combinations with batch selection strategies. In addition, we analyze effects of factors such as feature representation, database and seed size.

Chen et al. proposes instance-level and feature-level analysis of several active learning methods on word sense disambiguation [36]. In our paper, we take feature representation as a factor affecting the performance of active learning whereas Chen et al. uses active learning for feature selection. Thus, our work enlightens the missing side of the relationship of feature representation and active learning.

4.2 Active Scene Learning

Our framework distinguishes itself from the frameworks in the literature in terms of the segmentation aspect, properties of the training instances and multi-class labeling capability for the scenes. We will first compare our framework with the approaches targeting segmentation. Then, we will introduce several methods employing "bags approach" and point out the distinctions of our framework from these methods. Also we will present works utilizing structured instances and emphasize the distintion on the definition of segmentation.

Vezhnevets et al., Veeraraghavan et al., Ma et al. and Yan et al. apply active learning on image segmentation such that pixels are the primitives and they request labels for either pixels or regions of pixels [21, 22, 24, 37, 38]. Similarly, Pan et al. request labels for regions of pixels, but they use fixation-based active learning [39]. Han et al. use active learning for skin segmentation on video sequences and request labels for pixels [40]. Iglesias et al. apply active learning on CT scans and compute how informative a scan is by using estimated probabilities for each voxel in the scan [23]. Top et al. request labels for a plane of pixels in 3D image segmentation [41]. Tuia et al. using hierarchical clustering and apply active learning with a prune-and-label strategy [25]. All of these methods use pixels (or super pixels) as the training

sample whereas our framework requires a labeled object, which consists of a group of primitives and is extracted from a completely segmented scene, as the training sample. In other words, what we add as sample is an object itself, not a part of an object.

Vijayanarasimhan et al. and Xu et al. use multi-instance active learning for image categorization task [27, 29]. Li and Yeung apply multi-instance active learning for drug activity prediction, protein sequence classification and image retrieval [31]. Settles et al. and Zhou et al. employ active learning on content-based image retrieval and text classification [2, 30, 28]. These authors use the bags of data approach such that bags contain either segments (extracted via over segmentation) or objects. They label bags as positive or negative depending on whether the bag contains intended class or not. Rather than bags we use the term "scene" which contains one or more objects from various classes. Therefore, we do not give positive or negative labels to a scene in our framework. Our framework aims to select the most informative scenes or the most informative parts of the scenes which contain the objects that will increase the segmentation performance the most. In addition, these objects may belong to various classes rather than a single class.

Roth and Small propose active learning methods for structured output spaces on domains like semantic role labeling [32]. Settles and Craven propose various active learning methods, such as token entropy and sequence entropy, for sequence labeling task [7]. This task seems similar to sketch segmentation task, but our primitives (fragments) have no meaning by themselves, whereas tokens can have a meaning by themselves and labeling tokens can yield segmentation implicitly. However, fragments must be considered in groups (segments).

Chapter 5

FUTURE WORK

We see three main directions for future work: investigating more factors affecting the performance in detail, cost-sensitive analysis, and optimization of active learning systems.

5.1 Factor based Analysis

Although our empirical analysis employs a factor analysis on the performance of active learning, set of factors to investigate can be extended. One important factor besides the ones we analyzed can be the "batch size". It is the number of instances added to the training set per round. We set it to a small number according to our available resources. We believe that it is small enough for not to cause a bias among the batch selection strategies we applied. However, it would be another contribution to investigate the performance change of the experimented methods with various choices of batch size. We also suggest that batch size should be dynamic. Distribution of the informativeness scores, consecutive validation accuracies of the model, and a similarity metric fine tuned by considering relative distances of support vectors (for SVMs) might allow us to determine a dynamic batch size.

We analyzed seed set size as a factor affecting the active learning performance in our empirical analysis. We observed its affect the methods of single classifier based approach. We got the intuition that an initial model should pass an accuracy limit before active learning rounds start. However, it is not clear how to define such a limit. Future research should consider the initialization procedure as a dynamic procedure rather than providing a fixed size of seed set. "Check rounds" might be tried. For example, we can initilize the model with one instance from each class. At this point,

we can apply both an active learning round and a random learning round with the same configuration of the data and the same model. Depending on the performance difference on validation accuracy, we might decide to add more samples. If sources allow, such a strategy might be employed several times. When the source limit is passed, we can make a decision for using random learning or active learning.

Database is one of the factors in our analysis. Since we know some propeties of our databases (such as noiseness and style-variations), we could reach guiding intuitions. We can extend our work by experimenting on more databases. In addition, we can include a factor such that categorizes databases depending on the task. In this way, we might be able to question how properties of tasks and their specific data affect the performance. Besides, sparseness of data can be investigated as Sun and Hardoon [42]. In addition, we can try to investigate sparseness of instances from specific classes or clusters in the data.

5.2 Cost-Sensitive Analysis

In this thesis, we analyzed the performance by obtaining deficiency values of active learning methods with respect to random learning. Therefore, we consider the achievable accuracy values with labeling less data. A nice extention to our work might be the cost sensitive analysis of the active learning methods. Hence, cost can be defined in terms of the combination of time and effort/money spent for labeling an instance and the amount of labeled instances. By combining intelligent user interface designs and labeling schemes, we can reduce the cost even further.

We can suggest several labeling schemes for active scene learning. A strategy requesting the correction of all mis-segmentation and mis-labels might be the base line. We can improve such a system by providing corrections one by one, and we can update the segmentation and labeling after each correction is done. Also, we might not even need to provide the labels, but correcting the segmentation would be sufficient. There can be a variety of such schemes. Implementing and analyzing performances of these schemes will be a valuable step in active learning.

Sketch based (and/or pen based) interfaces will enable efficient designs not only for sketch recognition but also for the other domains. For example, we can correct segmentation of an object by just circling it. This can be applied for any segmentation task. We may provide labels by dragging labels and also use options such as tapping in order to label multiple instances (or parts of instances) at once. In addition, using gesture-like sketches can further reduce labeling cost for tasks such as sequence labeling and text categorization.

5.3 Optimization of Active Learning Process

We need active learning to quickly examine the data and to re-train the model as fast as possible. Since re-training of some models takes huge time when the number of training samples are high, user might wait a lot for new unlabeled data to be suggested for labeling. Such a scenario will be costly in terms of time. Therefore, we need to optimize and parallelize the active learning process as well as the model training procedure.

It is easy to handle optimization of the classification part of the process. Unlabeled data in the pool can be classified faster by employing parallel programming. The bottleneck is generally the re-training procedure. Since some machine learning algorithms (e.g. SVM) have a quadratic complexity, it takes more and more time to train such models as the number of training samples increases. Parallizing such algorithms might not be straightforward. However implementing them in parallel or adapting them for incremental training will be an important step for applicability of active learning [43, 44]. We should also adapt the user interfaces for distributed labeling. Such labeling applications should consider synchronization for presenting unlabeled data to annotators and merging the labeled data.

Chapter 6

CONCLUSION

We explained our work under two main chapters of this thesis. The empirical analysis we conducted on basic active learning methods is shared in Chapter 2 and the new active learning framework and approaches are shared in Chapter 3. This chapter includes the contributions brought by our study explained in these chapters. We will conclude this chapter with an overall discussion over our work.

6.1 Contributions

We present an extensive empirical analysis on active learning methods. Also we introduce a new active learning framework, the so-called "active scene learning" framework and clearly show its distintion from existing frameworks and its benefits for the active learning field. Moreover, we introduce several selection schemes under our framework.

A number of contributions make our work unique:

- This work is the first comprehensive study on the use of active learning for sketch recognition.

- We present results from an extensive analysis which shows that the utility of active learning depends on a number of practical factors that require careful consideration. Our factor analysis is examplary for future active learning researchers to investigate performance of active learning in a comrehensive manner.

- We show that the Margin based selection consistently outperforms other informativeness measures for isolated sketch recognition.

- We demonstrate that utilizing more seeds yields significantly better active learning performance for the single classifier based approaches for isolated sketch recognition.

- We show that the use of active learning brings definitive advantages in challenging databases when accompanied with powerful feature representations.

- We propose a new active learning framework which distinguishes itself from the previous frameworks in terms of the segmentation aspect, properties of the training instances and the allowed selection and labeling schemes.

- We introduce two selection schemes: scene-wise and segment-wise. Scene-wise selection methods are adaptation of the basic methods used in previous frameworks whereas segment-wise selection is more specific to our framework. We demonstarate the effective performance of both scene-wise and segment-wise selection schemes over random selection via a statistical analysis. Moreover, our analysis shows that our segment-wise selection scheme outperforms scene-wise selection schemes.

6.2 Discussion

We believe that our empirical analysis of active learning methods will serve as a practical guide for the future active learning users working on sketch recognition. The results of our work provides useful insights not only for sketch recognition but for the entire active learning community. Moreover, our factor based analysis on active learning performance will be examplary for future researchers to approach performance analysis in an elaborate manner. By such analyses, we will be able to learn the factors affecting the active learning performance and how they affect. Hence, we will have more aware control on the active learning process.

Our active scene learning framework targets enabling active learning for instances composed of full-objects from various classes and tasks including segmentation such

that grouping primitives into separate objects in a collection of primitives is the process of segmentation and object-level labeling of the groups of primitives is the process of labeling. We introduce two selection schemes which allow users to select the instances scene-wise or segment-wise. To investigate performances of active learning methods that we introduced under these schemes, we conducted experiements on non-interspersed sketches. We obtained promising results for our selection methods. These methods can directly be applied to tasks utilizing sequential data. Also our methods can be applied to tasks where data might not be sequential but the definition of segmentation holds such as interspersed sketch recognition. However, empirically analyzing our methods on such tasks will be a nice extension for our work.

A potential improvement on our work is to provide a cost-sensitive analysis. This analysis may include two parts. One part may be several simulations on various labeling schemes. For example, we might not need to correct all mis-segmentations and mis-labelings in a scene. We might provide one correction, and then ask the system to re-segment and re-classify the remaining part of the scene. Such a labeling scheme should yield a huge gain on labeling effort. However, we need real-world user studies as a second part of a cost-sensitive analysis. The bottleneck on the cost would be re-training the system. Therefore, paralellization of both the machine learning algorithms and the user interface for multi-annotator case would be complementary for our work.

BIBLIOGRAPHY

[1] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," in *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

[2] B. Settles, "From theories to queries: Active learning in practice," in *Workshop on Active Learning and Experimental Design (in conjunction with AISTATS 2010)*, vol. 16 of *Journal of Machine Learning Research: Workshop and Conference Proceedings*, pp. 1–18, 2011.

[3] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: an evaluation," *Machine Learning*, vol. 68, no. 3, pp. 235–265, 2007.

[4] C. Gasperin, "Active learning for anaphora resolution," in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, (Boulder, Colorado), pp. 1–8, Association for Computational Linguistics, June 2009.

[5] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *Advances in Neural Information Processing Systems 20* (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), pp. 593–600, Cambridge, MA: MIT Press, 2008.

[6] B. Settles, "Active learning literature survey," tech. rep., 2010.

[7] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *EMNLP*, pp. 1070–1079, ACL, 2008.

[8] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," 1994.

[9] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," Technical Report T2009-06, Swedish Institute of Computer Science, 2009.

[10] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in *Proceedings of the ACL*, pp. 183–190, 1993.

[11] A. K. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proceedings of ICML-98, 15th International Conference on Machine Learning* (J. W. Shavlik, ed.), (Madison, US), pp. 350–358, Morgan Kaufmann Publishers, San Francisco, US, 1998.

[12] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. J. Mooney, "Active learning for probability estimation using jensen-shannon divergence," in *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings* (J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, eds.), vol. 3720 of *Lecture Notes in Computer Science*, pp. 268–279, Springer, 2005.

[13] S. P. Engelson and I. Dagan, "Minimizing manual annotation cost in supervised training from corpora," in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, (Santa Cruz, California, USA), pp. 319–326, Association for Computational Linguistics, June 1996.

[14] D. Shen, J. Zhang, J. Su, G. Zhou, and C. L. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain* (D. Scott, W. Daelemans, and M. A. Walker, eds.), pp. 589–596, ACL, 2004.

[15] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Machine Learning, Proceedings of the Twentieth International Con-*

ference (ICML 2003), August 21-24, 2003, Washington, DC, USA (T. Fawcett and N. Mishra, eds.), pp. 59–66, AAAI Press, 2003.

[16] D. of the Army, *F. Manual, 101-5-1, Operational Terms and Graphics*. Washington, DC, 1997.

[17] R. Niels, D. Willems, and L. Vuurpijl, "The NicIcon database of handwritten icons," in *11th International Conference on the Frontiers of Handwriting Recognition (ICFHR 2008)*, (Montreal, Canada), August 2008.

[18] A. Khotanzad and Y. Hong, "Invariant image recognition by zernike moments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 5, pp. 489 –497, 1990.

[19] T. Y. Ouyang and R. Davis, "A visual approach to sketched symbol recognition," in *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, (San Francisco, CA, USA), pp. 1463–1468, Morgan Kaufmann Publishers Inc., 2009.

[20] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *J. Mach. Learn. Res.*, vol. 5, pp. 255–291, Dec. 2004.

[21] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *CVPR*, pp. 3162–3169, IEEE, 2012.

[22] H. Veeraraghavan and J. V. Miller, "Active learning guided interactions for consistent image segmentation with reduced user interactions," in *ISBI*, pp. 1645–1648, IEEE, 2011.

[23] J. E. Iglesias, E. Konukoglu, A. Montillo, Z. Tu, and A. Criminisi, "Combining generative and discriminative models for semantic segmentation of CT scans via

active learning," in *Information Processing in Medical Imaging - 22nd International Conference, IPMI 2011, Kloster Irsee,Germany, July 3-8, 2011. Proceedings* (G. Székely and H. K. Hahn, eds.), vol. 6801 of *Lecture Notes in Computer Science*, pp. 25–36, Springer, 2011.

[24] A. Ma, N. V. Patel, M. Li, and I. K. Sethi, "Confidence based active learning for whole object image segmentation," in *Multimedia Content Representation, Classification and Security, International Workshop, MRCS 2006, Istanbul, Turkey, September 11-13, 2006, Proceedings* (B. Günsel, A. K. Jain, A. M. Tekalp, and B. Sankur, eds.), vol. 4105 of *Lecture Notes in Computer Science*, pp. 753–760, Springer, 2006.

[25] D. Tuia, J. Muñoz-Marí, and G. Camps-Valls, "Remote sensing image segmentation by active queries," *Pattern Recognition*, vol. 45, no. 6, pp. 2180–2192, 2012.

[26] A. Culotta and A. Mccallum, "Reducing labeling effort for structured prediction tasks," in *In AAAI-05*, pp. 746–751, 2005.

[27] S. Vijayanarasimhan and K. Grauman, "Multi-level active prediction of useful image annotations for recognition," in *NIPS* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 1705–1712, Curran Associates, Inc, 2008.

[28] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," *CoRR*, vol. abs/0807.1997, 2008.

[29] X. Y. Xu and B. X. Li, "Multiple class multiple-instance learning and its application to image categorization," *International Journal of Image and Graphics*, vol. 7, pp. 427–444, July 2007.

[30] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *NIPS*

(J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), Curran Associates, Inc, 2007.

[31] W.-J. Li and D.-Y. Yeung, "MILD: Multiple-instance learning via disambiguation," *IEEE Trans. Knowl. Data Eng*, vol. 22, no. 1, pp. 76–89, 2010.

[32] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *Proceedings of the 17th European conference on Machine Learning*, ECML'06, (Berlin, Heidelberg), pp. 413–424, Springer-Verlag, 2006.

[33] R. S. Tumen and M. Sezgin, "Dpfrag: A trainable stroke fragmentation framework based on dynamic programming," *IEEE Computer Graphics and Applications*, vol. 99, no. PrePrints, p. 1, 2012.

[34] R. Arandjelovic and T. M. Sezgin, "Sketch recognition by fusion of temporal and image-based features," *Pattern Recognition*, vol. 44, no. 6, pp. 1225–1234, 2011.

[35] T. J. Markowitz, "An empirical evaluation of active learning and selective sampling variations supporting large corpus labeling," 2011.

[36] J. Chen, A. I. Schein, L. H. Ungar, and M. Palmer, "An empirical study of the behavior of active learning for word sense disambiguation," in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA* (R. C. Moore, J. A. Bilmes, J. Chu-Carroll, and M. Sanderson, eds.), The Association for Computational Linguistics, 2006.

[37] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *CVPR*, pp. 845–852, IEEE, 2012.

[38] C. Yan, D. Wang, S. Shan, and X. Chen, "Interactive segmentation with recommendation of most informative regions," in *IScIDE* (J. Yang, F. Fang, and

C. Sun, eds.), vol. 7751 of *Lecture Notes in Computer Science*, pp. 483–490, Springer, 2012.

[39] C. Pan, D. S. Park, H. Lu, and X. Wu, "Color image segmentation by fixation-based active learning with ELM," *Soft Comput*, vol. 16, no. 9, pp. 1569–1584, 2012.

[40] J. Han, G. M. Award, A. Sutherland, and H. Wu, "Automatic skin segmentation for gesture recognition combining region and support vector machine active learning," in *FG*, pp. 237–242, 2006.

[41] A. Top, G. Hamarneh, and R. Abugharbieh, "Active learning for interactive 3D image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2011 - 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III* (G. Fichtinger, A. L. Martel, and T. M. Peters, eds.), vol. 6893 of *Lecture Notes in Computer Science*, pp. 603–610, Springer, 2011.

[42] S. Sun and D. R. Hardoon, "Active learning with extremely sparse labeled examples," *Neurocomputing*, vol. 73, no. 16-18, pp. 2980–2988, 2010.

[43] E. Y. Chang, K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu, and H. Cui, "Psvm: Parallelizing support vector machines on distributed computers," in *NIPS*, 2007.

[44] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *NIPS* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), pp. 409–415, MIT Press, 2000.