Computers & Graphics 52 (2015) 93–105

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

Active learning for sketch recognition ☆

Erelcan Yanık *, Tevfik Metin Sezgin

Koç University, College of Engineering, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
30 July 2015
Accepted 30 July 2015
Available online 7 August 2015

Keywords:
Active learning
Sketch recognition
Empirical analysis
Factor analysis
ANOVA

ABSTRACT

The increasing availability of pen-based tablets, and pen-based interfaces opened the avenue for computer graphics applications that can utilize sketch recognition technologies for natural interaction. This has led to an increasing interest in sketch recognition algorithms within the computer graphics community. However, a key problem getting in the way of building accurate sketch recognizers has been the necessity of creating large amounts of annotated training data. Several authors have attempted to address this issue by creating synthetic data, or by building easy-to-use annotation tools. In this paper, we take a different approach, and demonstrate that the active learning technology can be used to reduce the amount of manual annotation required to achieve a target recognition accuracy. In particular, we show that by annotating few, but carefully selected examples, we can surpass accuracies achievable with equal number of arbitrarily selected examples. This work is the first comprehensive study on the use of active learning for sketch recognition. We present results of extensive analyses and show that the utility of active learning depends on a number of practical factors that require careful consideration. These factors include the choices of informativeness measures, batch selection strategies, seed size, and domain-specific factors such as feature representation and the choice of database. Our results imply that the Margin based informativeness measure consistently outperforms other measures. We also show that active learning brings definitive advantages in challenging databases when accompanied with powerful feature representations.

1. Introduction

Sketch recognition is an enabling technology that lies at the foundation of many computer graphics applications, including educational applications [1,2], graphics applications for design [3–5], shape retrieval [6], and animation [7]. A widely acknowledged problem in building accurate sketch recognition systems is the labor-intensive nature of obtaining large amounts of labeled data [8]. In this paper, we demonstrate the utility of the active learning technology in reducing the amount of manual annotation required to achieve a target recognition accuracy. The results and the approach presented in this paper provide valuable insights to the practitioners of sketch recognition as well as the broader community of computer graphics practitioners who rely on machine learning in their applications.

The sketch recognition community has attempted to address the data labeling problem by synthesizing artificial training examples from few labeled examples [8], by building custom interfaces for labeling data [9–12] or by applying automated labeling supported with a partially trained recognizer [13]. Each case requires human annotators to label data without being particularly selective about which samples are labeled. We show that, using active learning, it is possible to prioritize the labeling process in a way that allows one to build more accurate classifiers with fewer labeled instances, hence reduce the annotation 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. The basic premise of active learning is that some training examples carry more information than others. Hence, if we can identify them among the unlabeled examples, and have them labeled by a human annotator, we can potentially converge to higher accuracies with substantially less human annotation effort.

Active learning process is initialized by training a classifier with a few labeled samples, the so-called "seed set". The learning process continues in rounds until a target validation accuracy is achieved or until we run out of resources (e.g. time or computational resources). In each round, we train a classifier with the available labeled data, and use it to classify the unlabeled examples. We then use the scores assigned to the unlabeled samples to

☆This article was recommended for publication by Beryl Plimmer

* Corresponding author.
E-mail addresses: eyanik13@alm.ku.edu.tr (E. Yanık), mtsezgin@ku.edu.tr (T.M. Sezgin).
URL: http://www.iui.ku.edu.tr (T.M. Sezgin).

http://dx.doi.org/10.1016/j.cag.2015.07.023
0097-8493/

select the subset of most useful samples for subsequent labeling. The round ends by adding the newly labeled data to the training set and re-training the model.

Despite its theoretical appeal, recent empirical results show that active learning does not always yield the expected benefits in practical real world problem settings [14]. For example, Schein and Ungar report inconsistent and negative results for active learning [15]. Likewise, Gasperin reports that none of the experimented active learning methods reached a remarkable performance although they converge on different sets of training examples from each other [16]. Guo and Schuurmans also point out that active learning methods perform poorly with respect to random learning, which is the strategy of selecting samples randomly from a pool of unlabeled examples [17]. Therefore, there is a practical and real need for analyzing the empirical performance of active learning in various settings in order to understand if active learning is effective at all, and if so under which conditions. In this paper, we present such an analysis for the sketch recognition domain to identify the set of practical issues one should be aware of when using active learning, and investigate how these issues affect active learning performance.

Specifically, we investigate the performance of active learning under combinations of a large variety of informativeness measures and batch selection strategies, as well as factors such as feature representation, database and seed set size for sketch recognition. Our analysis results constitute a detailed and practical guide for active learning users for sketch recognition and provide valuable insights for machine learning practitioners in the computer graphics community. Our main contributions can be summarized as follows:

- We present a set of carefully designed experiments and a battery of accompanying statistical tests, which will serve as a roadmap to follow for practitioners of active learning who wish to perform factor analysis.
- We present the first extensive empirical analysis on active learning for sketch recognition, and provide a detailed discussion of the analysis results.
- We determine the best performing and reliable informativeness measure for sketch recognition.
- We show that starting with a large seed set yields better active learning performance for the single classifier approach.
- We show that the use of active learning brings definitive advantages in challenging databases when accompanied with powerful feature representations.

This paper 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 measure employed in our analysis and then present the analysis methodology. We present the analysis results with a discussion in Section 5. Finally, we conclude with related work and a summary of future research directions.

2. Active learning methods

There are two essential steps in active learning: measuring informativeness of unlabeled samples and selecting batches of collectively informative samples which are mutually non-redundant. In this section, we describe informativeness measures and batch selection strategies that are used in our experiments.

2.1. Informativeness measures

There are two main approaches for measuring informativeness: the *single classifier* approach and the *query by committee* (QBC) approach. Measures of informativeness are based on the rationale that samples that a classifier cannot confidently classify, or a group of classifiers disagree on can potentially supply more information when labeled. We list the informativeness measures included in our experiments in Table 1. Four of the measures follow the single classifier approach, in which decisions are based on a single classifier's prediction on a sample. The other four measures follow the query by committee (QBC) approach, in which the disagreement of the committee members on the label of a sample is used to derive informativeness.

2.2. Batch selection strategies

Active learning requires classifiers to be retrained as more labeled data gets added to the training set. Since training is costly, newly labeled examples are usually added in batches, rather than one by one. Although adding samples in batches reduces computational requirements, it bears the risk of adding samples which carry mutually redundant information. In particular, two samples which are extremely informative when taken individually may actually contain similar and redundant information, so including them both in the training data may not yield extra advantage over having just one or the other. Hence, we should avoid sets containing mutually redundant samples. Several batch selection strategies have been proposed in the literature to avoid this problem, and we included four of them in our experiments.

Our empirical analysis includes the following batch selection strategies: Default selection, Global-FV strategy, Global-PE strategy

Table 1
Brief description of the informativeness measures used in our analysis.

	Informativeness measures	A sample is considered informative when:
Single classifer approach	Entropy based Selection[a] [18,19]	The entropy is high on class probabilities of a sample.
	Least Confident based Selection [18,20]	The most likely class probability of a sample has a low value.
	Margin based Selection[a] [18]	The difference of the most and the second most likely class probabilities of a sample has a low value.
	Körner–Wrobel Selection[a] [21]	The Körner–Wrobel value computed for the sample is low. It is a combination of Least Confident and Margin based selection strategies.
Query by committee approach	Kullback Leibler Divergence based Selection [21,22]	KL-Divergence among the committee on a sample is high.
	Jensen Shannon Divergence based Selection [21,23]	JS-Divergence among the committee on a sample is high.
	Vote Entropy based Selection [21,24]	The entropy of the class label votes of the committee is high.
	Weighted Vote Entropy based Selection [21]	The weighted entropy of the class label votes of the committee is high.

[a] The method has implementation also for the query by committee approach, in the literature, but we only include the single classifier version.

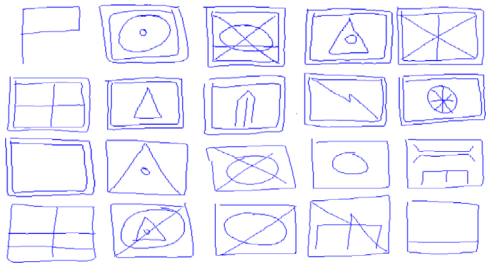

Fig. 1. Example sketches from each class of COAD database.

Fig. 2. Example sketches from each class of NicIcon database.

and Combined strategy. N samples with top informativeness scores are selected in the Default batch selection strategy, where N is the batch size. Global-FV and Global-PE are based on clustering, similar to the method described by Shen et al. [25]. We select N samples with top informativeness scores, divide them into K clusters, and add the cluster centers to the batch, where K is the batch size. We set N to R*NumberOfClasses, where R is fixed to a small number such as 3 in order to avoid samples with very low informativeness score. Global-FV uses feature vectors for clustering whereas Global-PE uses probability estimates. Combined strategy is implemented as described in Brinker's paper [26] such that equally weighted informativeness and diversity (dissimilarity from samples currently in batch) scores are added to compute a final score for an unlabeled sample.

3. Experimental design

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

3.1. Databases and feature representations

We conducted our experiments on two databases containing hand drawn sketches of domain symbols. The first database is the Course of Action Diagrams (COAD) database [27], and the other database is the publicly available NicIcon database [28]. The COAD database contains a total of 620 samples from 20 different symbol classes, whereas the NicIcon database contains a total of 22,958 samples from 14 different classes. We present example sketches from each database for each class in Figs. 1 and 2.

The NicIcon database consists of sketches which were collected in 3 different sizes (small, medium, large) from 32 participants whereas the COAD database is collected from 8 participants without any specific size requirements. Therefore the NicIcon database contains more variation in style compared to the COAD database. In addition, the NicIcon database contains more noise compared to the COAD database probably due to fatigue of the participants in the data collection process. Moreover, the empirical results in the literature suggest that much higher accuracies can be achieved on the COAD database compared to the NicIcon database for various choices of features [27,29]. Hence, the NicIcon database is considered to be a more challenging database. Including these two datasets in our experiments allows us to investigate the performance of active learning on a harder database in comparison to an easier one in terms of style variation and noise.

For feature extraction, we used two image-based methods: Zernike Moments [30] and IDM [31]. Note that IDM features are considered to be more effective compared to Zernike Moments as empirical results in the literature suggests [29,31].

In active learning, seed size refers to the number of labeled examples used to bootstrap the learning process. In order to measure the effect of seed size on the active learning performance, we experimented with two seed size values: 1 labeled sample per class (small seed size), 4 labeled samples per class (large seed size). For each of the seed size choices, and each database (NicIcon/COAD), we created 10 *randomized starting sets*. This results in a total of 40 randomized starting sets as shown in Fig. 3. The randomized starting sets serve as unique initial starting conditions, which we try to improve upon using active learners equipped with various learning strategies.

3.2. Trials

A trial refers to the end-to-end process of active learning (or random learning) on a randomized starting set, throughout which we measure the classification accuracy of the model for each round. We initialize each trial by training the classifier or the committee members with the seed set. Then, the process continues by selecting and adding 10 samples to the training set and re-training the model at each round. Each trial continues until all the unlabeled data in the training pool is labeled. Since the NicIcon database has a fairly large training pool, we limited the number of rounds for this database to 120. This limit is sufficient for the test accuracy of the classifier to saturate.

We conducted trials for all combinations of 2 feature representations, 4 batch selection strategies and 8 informativeness measures for each randomized starting set. In addition, we carried out trials for random selection strategy for each choice of feature representations for each randomized starting set. Therefore, we conducted a total of 66 trials for each randomized starting set as demonstrated in Fig. 4. In total, we conducted 2640 trials.

3.3. Classifier design

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

4. Analysis

4.1. Deficiency measure

In order to assess the relative performance of various active learning methods, we used the deficiency measure described by Baram et al. [32]. The deficiency of method A with respect to B, deficiency(A,B), is a standard measure of the relative performance of algorithms throughout the active learning process. Fig. 5 demonstrates computation of the deficiency value given the maximum accuracy line, and the accuracy curves of two learning methods.

The maximum accuracy line represents the accuracy of the classifier when it is trained with all the training data in the pool. An accuracy curve represents the list of accuracies (over the test set) achieved in each round of the active learning process after the

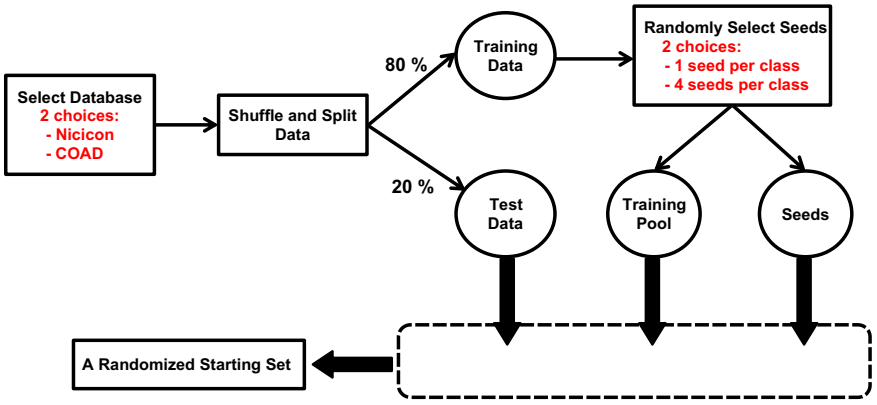

Fig. 3. The flow chart describing the process of creating a randomized starting set for a set of parameters (i.e. the choices of database and seed set size). 2 choices of database, 2 choices of seed set size and 10 shuffles of the selected dataset yield 40 randomized starting sets.

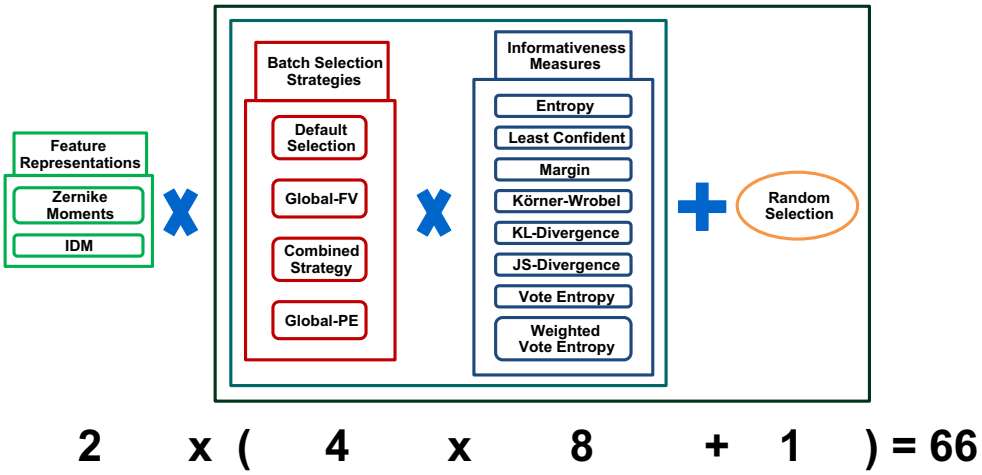

Fig. 4. For each randomized starting set, we train active learners with different feature representation, batch selection, and informativeness settings. In addition, we also train a random learner for each randomized starting set using each feature representation.

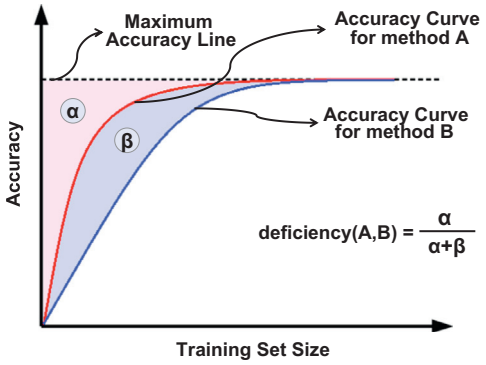

Fig. 5. The deficiency is defined as the ratio of the area between the accuracy curve of method A and the maximum accuracy line; and the area between the accuracy curve of method B and the maximum accuracy line.

classifier is trained with the available labeled data. Let $D = deficiency(A, B)$ be the deficiency of algorithm A computed with respect to algorithm B. $D = 1$ implies that the methods have a similar performance. Values less than one imply that method A is superior, while values greater than one imply that method B has superior performance.

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 multiway ANOVA tests. 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. In cases where sphericity is violated, the degrees of freedom have been corrected by the Greenhouse-Geisser correction. We also performed Levene's test to check the homogeneity of variances between groups and used transformed values where appropriate. Bonferroni corrected paired *t*-tests were performed as post hoc tests, in order to explore the mean differences across the levels of the concerned factors.

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

Table 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 (I)	$F(7,252)=515.287$	$p=0.000$
Batch selection strategy (BS)	$F(2.290,105.131)=31.565$	$p=0.000$
Feature 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 ∗ I	$F(21,756)=12.087$	$p=0.000$
SS ∗ I	$F(7,252)=13.060$	$p=0.000$
DB ∗ I	$F(7,252)=180.931$	$p=0.000$
DB ∗ FR ∗ I	$F(7,252)=69.861$	$p=0.000$

informativeness measure. The deficiency value was taken as dependent variable, which was computed for each active learner with respect to the random learner (values less than 1 indicate the active learning outperforming the random baseline).

To further observe the effect of seed size on the performance of active learners, we conducted a 4-way Mixed ANOVA. For this design, the between group variable is database; and within group variables are feature representation, batch selection strategy and informativeness measure. As the dependent variable, we computed deficiency values comparing active learners initialized with small and large seed sets. Hence, we directly compare these two cases rather than taking the performance with respect to random selection as a reference point. Note that the deficiency values were computed for accuracies obtained past the large seed set size limit. In other words, accuracies obtained with fewer samples than the large seed set size limit are not considered in order to have a fair comparison among two cases.

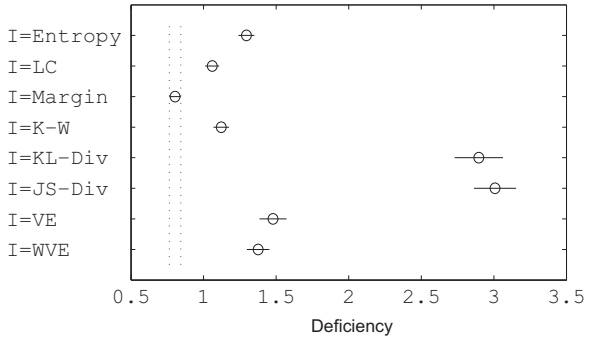

Fig. 6. The estimated marginal means for the informativeness measure factor. Margin-based selection is the only informativeness measure performing significantly better than random selection.

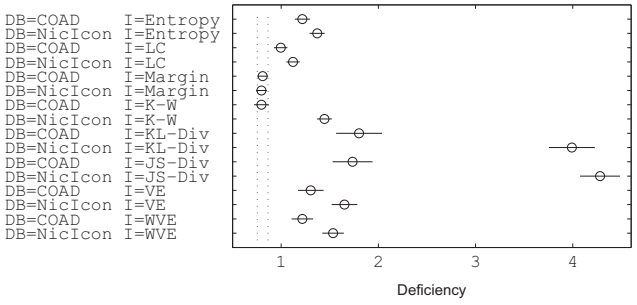

Fig. 7. The estimated marginal means for the 2-way interaction of informativeness measure and database factors. The significant advantage of Margin-based selection over random selection is consistent on two databases.

5. Results

In this section, we present the results of our factor analysis along with detailed discussion addressing issues such as the choice of informativeness measure, the choice of batch selection strategy, the effect of seed size and strategies for utilizing prior knowledge in active learning.

Throughout this section, we frequently resort to estimated marginal mean graphs for reporting test results. In an effort to assist the reader, we would like to present a practical guideline for interpreting these graphs:

- If confidence intervals of two marginal means do not intersect, this implies a statistically significant difference. Also, note that all confidence intervals in our analysis are 95% confidence intervals.
- If the performance difference between two methods is statistically significant, then the superior method is the one with smaller deficiency. Note that all performances (in 5-way ANOVA) are measured with respect to the performance of random selection since deficiency values are computed against the performance of random selection.
- A method performs confidently better than random selection if upper bound of its confidence interval is less than 1.

F-scores and *p*-values for 5-way Mixed ANOVA analysis are presented in Table 2. We will refer to this table throughout Section 5.

5.1. Choice of the informativeness measure

For a particular learning task in a given domain, selecting an informativeness measure is the first step of building an active learner. Although this is the most crucial step, there are no guidelines and general rules for selecting the “right” informativeness measure. Hence, when it comes to selecting an informativeness measure, empirical results obtained for a variety of feature extraction methods on representative databases serve as

Table 3
Bonferroni corrected paired *t*-test results for batch selection strategies. A mean difference less than zero indicates that the method performs significantly better (has a confidently smaller deficiency value) than the reference method. All batch selection strategies perform significantly better than Default selection whereas Global-FV strategy performs the best among all.

(I) Batch selection strategy	(J) Reference method	Mean difference (I-J)	Std. error	Sig.	95% confidence interval for difference	
					Lower bound	Upper bound
Default	Global-FV	0.299*	0.034	0.000	0.203	0.395
	Combined	0.090*	0.030	0.026	0.007	0.173
	Global-PE	0.122*	0.033	0.004	0.031	0.213
Global-FV	Default	−.299*	0.034	0.000	−0.395	−0.203
	Combined	−0.209*	0.030	0.000	−0.294	−0.124
	Global-PE	−0.177*	0.031	0.000	−0.264	−0.090
Combined	Default	−0.090*	0.030	0.026	−0.173	−0.007
	Global-FV	0.209*	0.030	0.000	0.124	0.294
	Global-PE	0.032	0.031	1.000	−0.055	0.119
Global-PE	Default	−0.122*	0.033	0.004	−0.213	−0.031
	Global-FV	0.177*	0.031	0.000	0.090	0.264
	Combined	−0.032	0.031	1.000	−0.119	0.055

indispensable sources of information. In this subsection, we will present our analysis results on the informativeness measure. Our analysis shows that, for our domain of interest, Margin-based informativeness measure consistently outperforms other measures by a large margin, irrespective of factors such as the choice of database and the feature extraction method.

Informativeness factor has a significant effect on active learning performance against random learning as shown in Table 2. Estimated marginal means for informativeness measures are presented in Fig. 6. Note that only Margin-based informativeness measure has an upper bound (on its confidence interval) less than one. Therefore, only Margin-based informativeness measure can perform confidently better than random selection. Also note that divergence based methods perform significantly worse than all the other informativeness measures.

As shown in Fig. 7, the superior performance of Margin-based informativeness measure is consistent across databases as well. In particular, it outperforms random selection in both databases. Moreover, observe that there is no significant difference in the performance of Margin-based informativeness across databases. Also observe that Körner–Wrobel informativeness measure performs as well as Margin-based informativeness measure on the COAD database, but it performs significantly worse than random selection in the NicIcon database. Hence, Margin-based selection stands out due to its consistency across databases as well as its superior performance.

5.2. Choice of the batch selection strategy

Due to the re-training penalty inflicted at each round of the active learning process, users tend to label and add samples to the training set in batches rather than adding one sample at a time. However, the naíve approach of grouping the individually most informative samples may yield batches containing mutually redundant instances. Therefore, an elaborate batch selection strategy is crucial for obtaining the desired benefit from active learning. In this section, we will present a comparison of the basic batch selection strategies paired with the best performing informativeness measure (Margin-based selection), along with an overall comparison of these strategies.

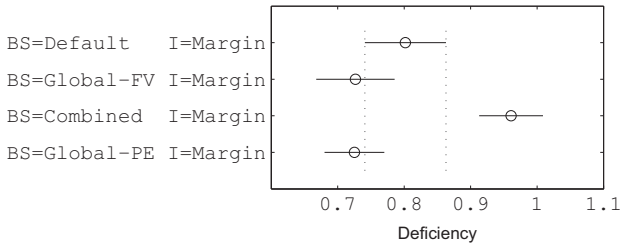

Fig. 8. The estimated marginal means for Margin-based selection over batch selection strategies. Margin-based selection can have a promising performance even without a sophisticated batch selection strategy.

It was shown in Table 2 that the batch selection strategy (BS) factor significantly affects active learning performance. We further investigated performances of the batch selection strategies via post hoc tests as presented in Table 3. As shown in the table, all batch selection strategies perform significantly better than the Default batch selection strategy. Also observe that Global-FV performs significantly better than all other methods. Therefore, we can conclude that for informativeness measures we considered in our experiments, Global-FV is the most appropriate batch selection strategy, whereas Default batch selection is the least desirable.

In the previous section, we reported that Margin-based selection is the only informativeness measure that can perform significantly better than random selection and its performance is consistent across databases. Hence, we investigated the performance of batch selection strategies specifically on Margin-based informativeness in addition to investigating the overall performance of the batch selection strategies. This allows us to observe how the batch selection strategies behave when coupled with an effective informativeness measure.

Results presented in Fig. 8 show that although coupling Margin-based informativeness with Global-FV and Global-PE strategies tends to yield a better mean performance, this superiority is not statistically significant compared to coupling with the Default batch selection strategy. Therefore, using Default batch selection with Margin-based informativeness yields satisfactory performance and also saves computation power, time, and implementation effort required by sophisticated batch selection strategies. Thus, we can couple Margin-based informativeness with Default batch selection strategy (rather than Global-FV or Global-PE) if we have scarce resources (time and computation power).

Another observation from Fig. 8 is that, the Combined strategy performs significantly worse than the other batch selection strategies when coupled with Margin-based selection. Note that Combined strategy uses a linear combination of the informativeness and diversity scores to score an unlabeled sample. In addition, the Combined strategy weights the informativeness and diversity scores equally. In this respect, the combined strategy can be regarded as an extension of the Default batch selection strategy that takes diversity into account in addition to informativeness. Hence, incorporating a measure of diversity into the batch selection measure appears to result in a performance deterioration. However, it is not clear if this is due to a genuine deficiency on the part of the diversity metric, or due to the inappropriateness of the arbitrary equal weighting scheme.

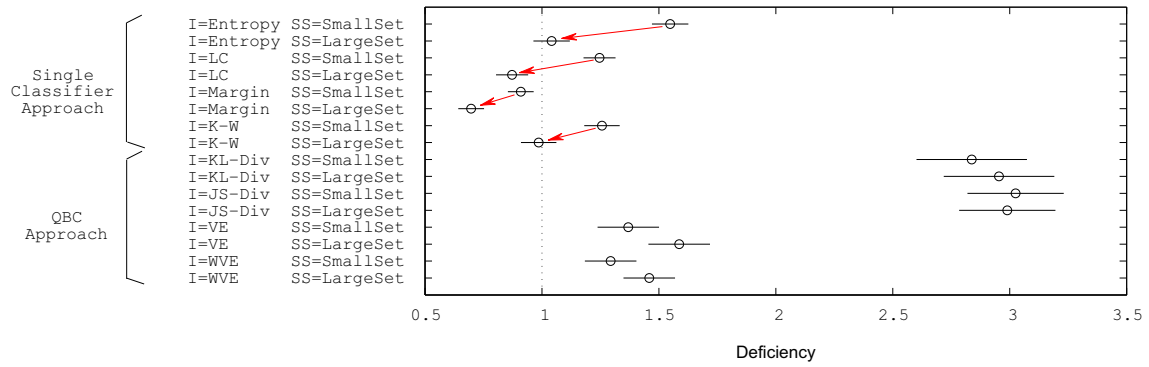

Fig. 9. The estimated marginal means for the 2-way interaction of informativeness measure and seed set size factors. All single classifier based methods perform significantly better when a larger seed set is utilized.

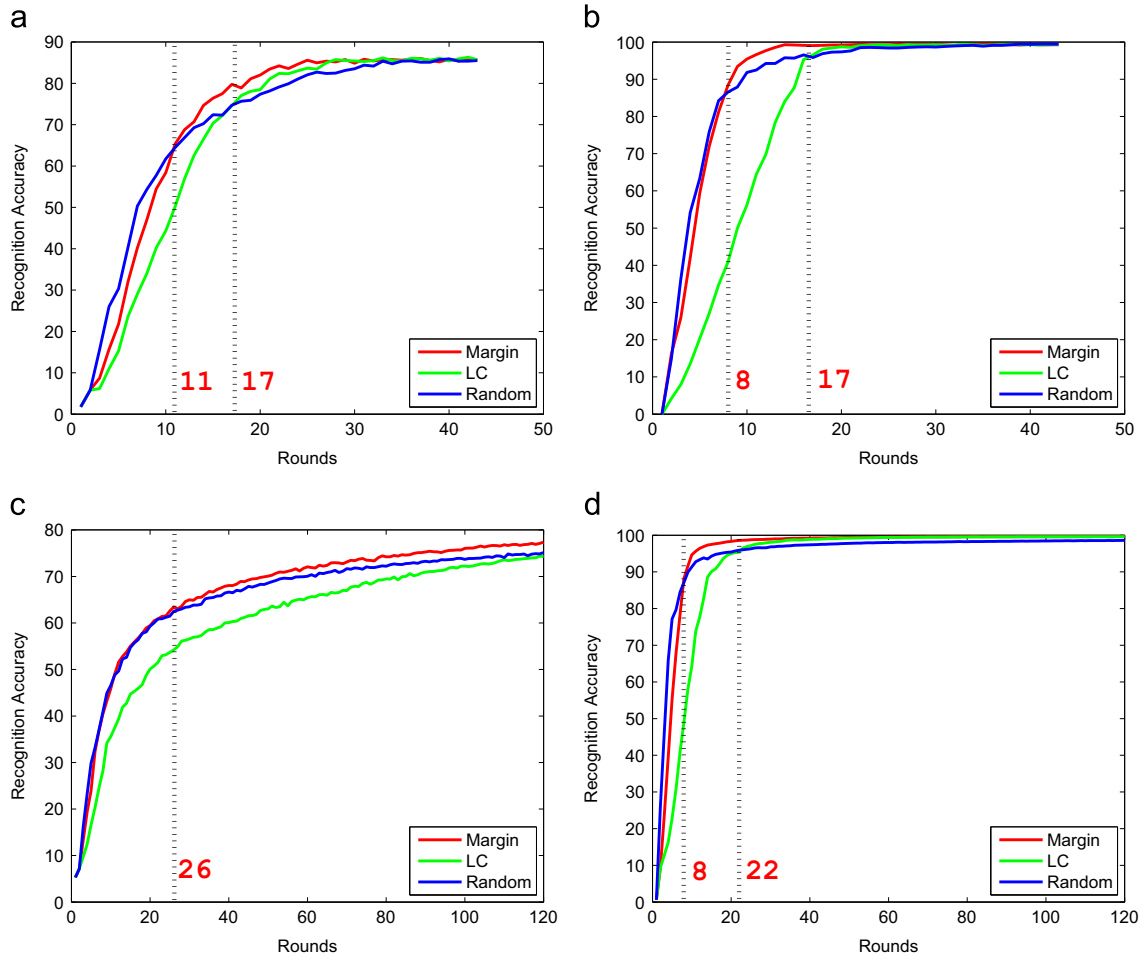

Fig. 10. Mean accuracy curves for single classifier based informativeness measures and random selection for each combination of databases and feature representations when small seed set is used. Dotted lines indicate when informativeness measure becomes beneficial with respect to random selection. (The numbers in red highlight the corresponding round numbers for the dotted lines.) (a) COAD-Zernike. (b) COAD-IDM. (c) NicIcon-Zernike. (d) NicIcon-IDM. (e) COAD-Zernike. (f) COAD-IDM. (g) NicIcon-Zernike. (h) NicIcon-IDM. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

5.3. *Effect of the seed size*

There is no feasible way of determining a globally optimal seed size, and seed size selection is considered to be an open problem in active learning. In an effort to come up with a seed size selection guideline, we carried out a detailed analysis investigating the effect of seed size on the performance of active learning.

It was shown in Table 2 that 2-way interaction of seed size and informativeness measure factors has a significant effect on the performance of active learning although the seed size (*SS*) factor does not. Further investigation shows that initializing active learners with larger seed sets yields significantly better results when coupled with single classifier based informativeness measures (see Fig. 9). This can be explained by further analysis of mean accuracy curves, which illustrate that using a large seed size yields a strong classifier early on, which in turn results in more accurate active learning decisions earlier in the process. Figs. 10 and 11 demonstrate that the single classifier based informativeness measures[1] make better decisions in earlier iterations and beat random selection much earlier when initialized with large seed set in comparison to small seed set. Hence, it is better to have as large of a seed set as resources allow.

To further investigate the effect of seed size, we conducted 4-way Mixed ANOVA as described in Section 4.2. Here, our goal is to check the relationship between the performance of informativeness measures and the choice of seed size. Hence, we compared deficiency values obtained with large and small seed set. This gives a direct comparison of performance for large seed sets and small seed sets rather than taking their relative performance with respect to random selection as a reference point. Therefore, we check whether improved decisions in sample selection due to large seed set can amortize the cost of additional labeling required to have a large seed set. Since the deficiency values in Table 4 represent the direct comparison between the cases utilizing large and small seed set, we can deduct that active learning yields higher accuracies with the same amounts of data when initialized with large seed set in comparison to small seed set.

Table 4 also points out that Margin-based informativeness is the least sensitive measure to the choice of seed size. In other words, it is the least affected measure by the choice of seed size. The upper bound of its confidence interval intersects with the deficiency value of one. Hence, for Margin-based selection there is no statistically significant difference across large and small seed sets although it tends to perform better with large seed set as its mean value suggests.

[1] We provide plots for each combination of database, seed set size, feature representation and single classifier based informativeness measures, combined with the Default batch selection strategy. Since plots are similar for other batch selection strategies, we omit these plots to save space and avoid clutter. Also, we do not include all informativeness measures in the same plot in order to have better visuality.

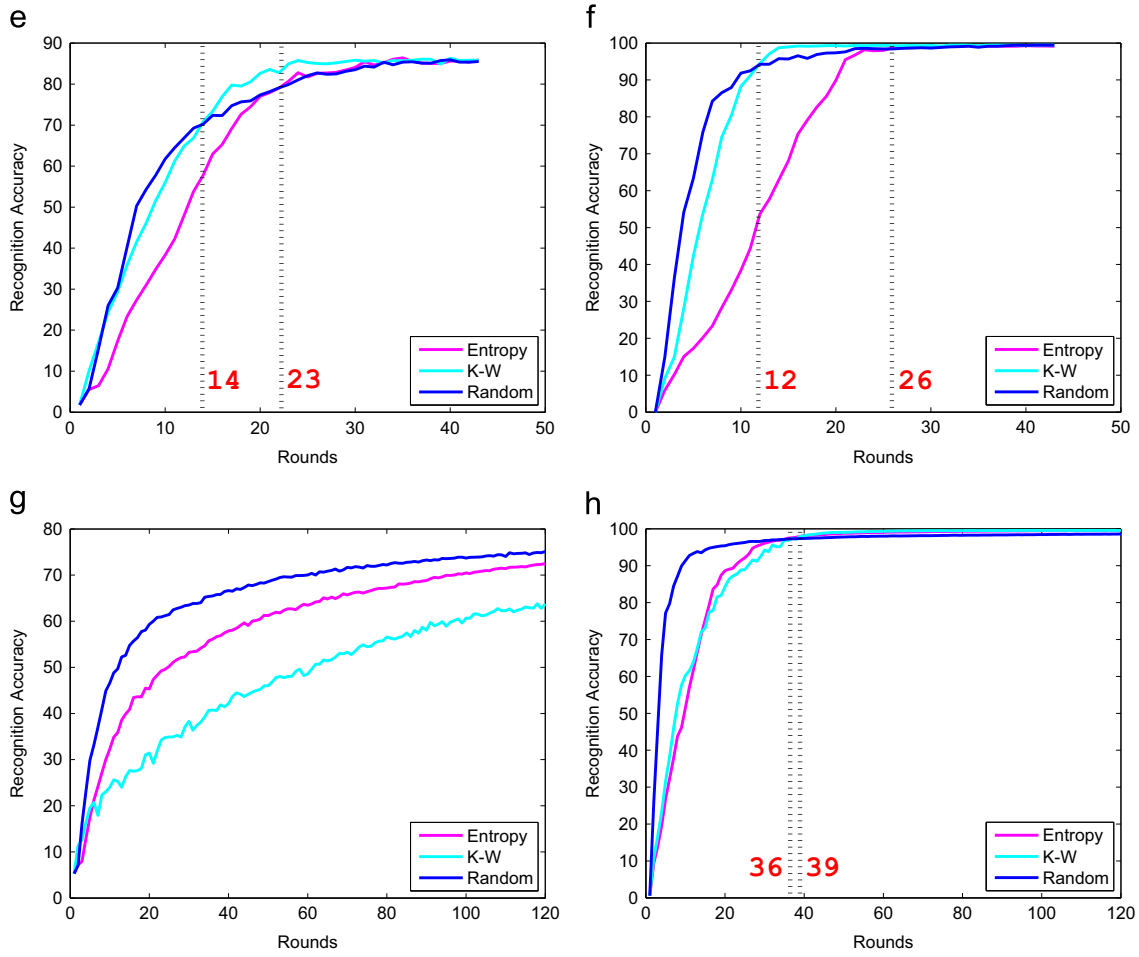

Fig. 10. (*continued*)

5.4. Utilizing prior knowledge

The NicIcon database is a more challenging database compared to the COAD database [27,29] and, the IDM features are considered to be more effective than Zernike moments [29,31], especially in more challenging databases. We ran a series of tests to understand if such prior knowledge on databases, or feature extraction methods can be used to guide the construction of the active learning setup. Our tests explored performance changes of active learners under two way interactions of the database and feature representation factors. Since only Margin-based informativeness can outperform random selection, we focus our discussion on Margin-based selection.

The most important result inferred from our experiments is that active learning brings definitive advantages in challenging databases when accompanied with powerful feature representations. This is depicted in Fig. 12 where the active learning performance is not significantly affected by the feature representation on the COAD database, while the performance on the NicIcon database is significantly higher with the IDM features compared to Zernike features, again as depicted in Fig. 12. Also, we observe that active learning proves to be ineffective when Zernike features are used with the NicIcon database. Hence, we recommend avoiding active learning if the feature representation is weak or unknown (e.g. a new feature representation scheme with no prior knowledge on its performance).

Another interpretation of these results suggests that samples under a strong feature representation are representative and informative by themselves on simpler databases. Therefore, random selection may work just as well. Since the COAD database is less noisy and has smaller style variations compared to the NicIcon database, its samples are more representative of the whole database compared to the NicIcon database. Hence, the active learning performance is *almost significantly lower* on the COAD database than the NicIcon database when IDM features are used (as shown in Fig. 12). In other words, it is harder to beat random selection on simpler databases when a strong feature representation is in use.

In conclusion, when the database is too simple and the feature representation is very strong; or when the database is too hard to learn and the feature representation is too weak, active learning may not yield the desired benefits. Thus, relying on active learning on such cases might be more costly than using random selection.

6. Discussion

6.1. Actual savings in the annotation effort

We based our statistical analysis on deficiency measure which is an established measure in the literature. To better demonstrate the labeling gains made by active learning, we also provide results demonstrating the number of samples to be labeled (including the seed set) in order to achieve the highest possible accuracy that both active learning and random selection can achieve. Table 5 demonstrates that active learning can make great savings in the number of samples to be labeled

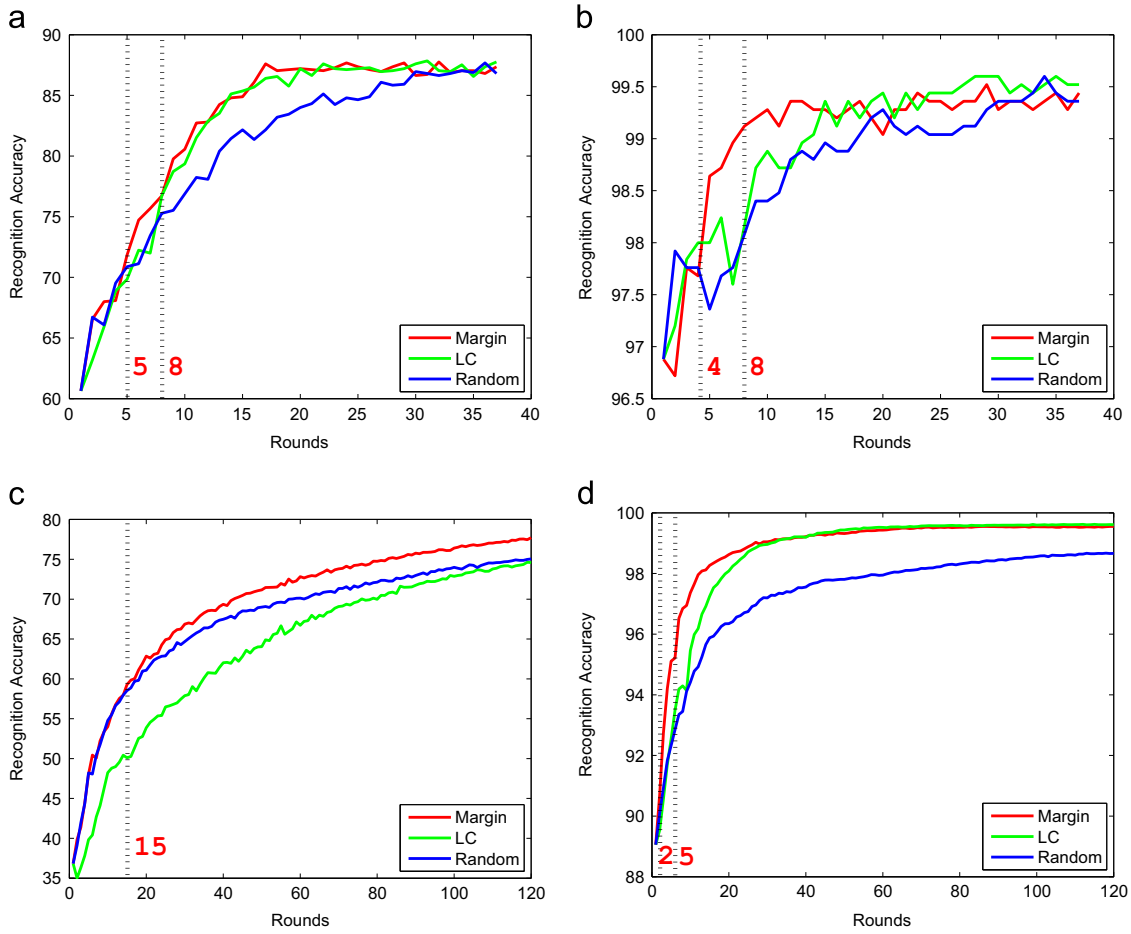

Fig. 11. Mean accuracy curves for single classifier based informativeness measures and random selection for each combination of databases and feature representations when large seed set is used. Dotted lines indicate when informativeness measure becomes beneficial with respect to random selection. (The numbers in red highlight the corresponding round numbers for the dotted lines.) (a) COAD-Zernike. (b) COAD-IDM. (c) NicIcon-Zernike. (d) NicIcon-IDM. (e) COAD-Zernike. (f) COAD-IDM. (g) NicIcon-Zernike. (h) NicIcon-IDM. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

compared to random selection. In addition, active learning can achieve % 99 accuracy on both databases by only labeling 0.26 of the training set of the COAD database and 0.012 of the training set of the NicIcon database.

6.2. Factors effecting active learning performance

Now, we turn our discussion on the factors and the possible reasons why they might have the effects reported in Section 5. We will present our reasoning on the effect of seed set size and utilizing prior knowledge, and then make additional comments on the other factors.

6.2.1. Effect of the seed size

We showed (in Section 5.3) that initializing active learning with as large of a seed set as resources allow yields better performance for single classifier based informativeness measures. Remember that informativeness values are computed over class probabilities of the samples which are assigned by the classifier at that moment (round). If active learning starts with an overly naïve classifier with an arbitrary decision boundary, the error on the class probabilities will be high. Therefore, informativeness values will be unreliable. By contrast, if the initial decision boundary is representative of the ultimate decision boundary, informativeness values will be more reliable. Therefore, we recommend to initialize active learning with large seed set as much as resources allow.

6.2.2. Effect of the database and the feature representation

For our empirical analysis, rather than testing on many arbitrarily picked databases, we carefully selected two databases. It is well known from the empirical results reported in the literature that the NicIcon database is more challenging database compared to the COAD database [27,29]. Moreover, the NicIcon database has more style variation and noise than the COAD database as we discussed in Section 3.1. We deliberately selected these two databases in order to examine both ends in terms of database difficulty. Also, empirical results on various databases suggest that IDM is a relatively powerful feature representation compared to Zernike Moments [29,31]. Hence, our analysis covers both ends for databases as well as feature representations by including these carefully and deliberately selected databases and feature representations. This allows us to investigate whether we can utilize prior knowledge about the complexity of the database and/or the strength of the feature representation for active learning.

The results in Section 5.4 demonstrate that when the feature representation is strong, fewer labeled data is sufficient to achieve high accuracies. Since each sample will carry more information due to better representation, any random subset of data tends to contain more information (see the random selection curves in Figs. 10 and

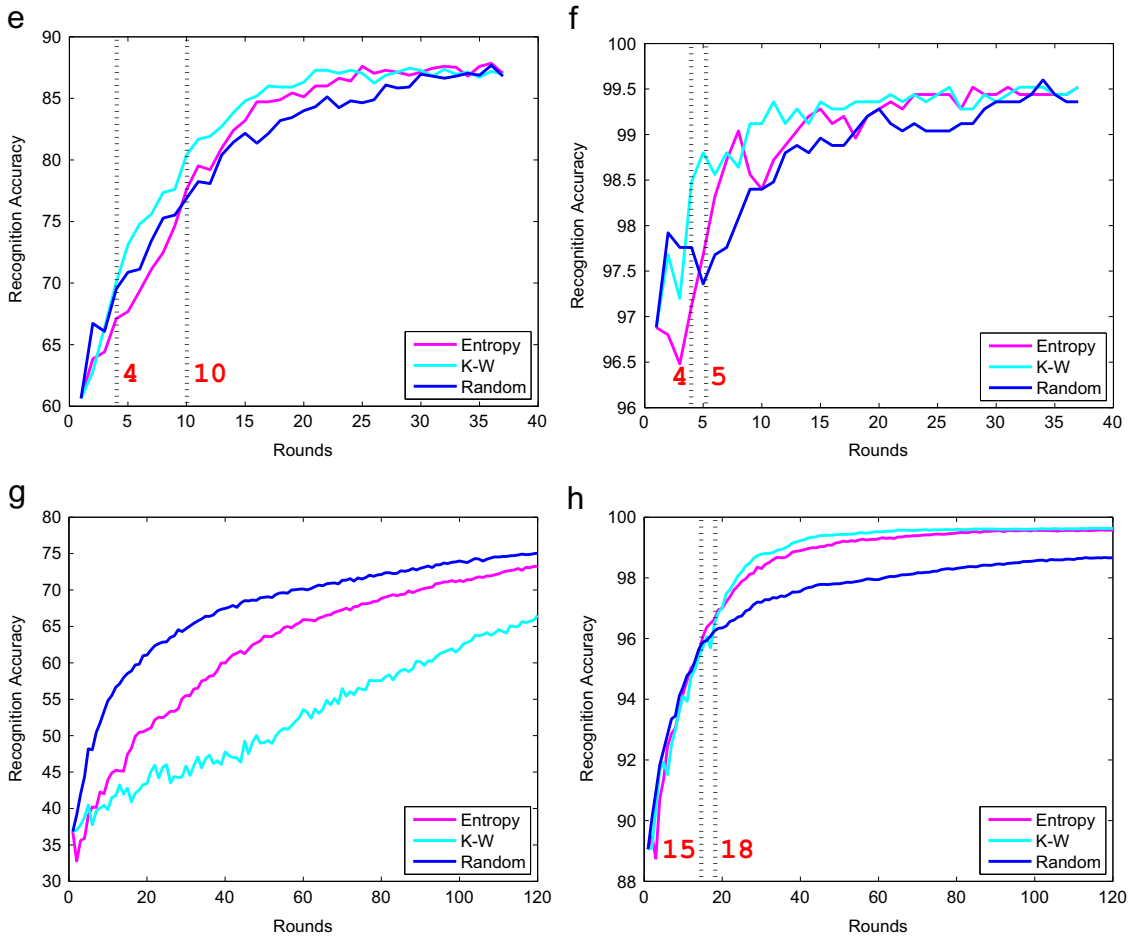

Fig. 11. (*continued*)

Table 4
For each informativeness measure, estimated marginal mean of the achieved deficiencies obtained using large seed set over small seed set presented as a result of the 4-way Mixed ANOVA analysis. 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
Single classifer approach	Entropy	0.489	0.022	0.442	0.536
	LC	0.630	0.064	0.495	0.765
	Margin	0.846	0.098	0.639	1.053
	K-W	0.784	0.102	0.570	0.997
Query by committee approach	KL-Div	0.478	0.027	0.421	0.535
	JS-Div	0.435	0.022	0.389	0.481
	VE	0.789	0.041	0.703	0.874
	WVE	0.785	0.055	0.668	0.901

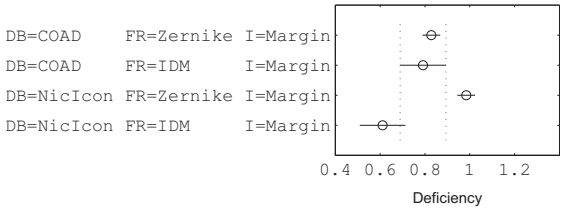

Fig. 12. The estimated marginal means for Margin-based selection over 2-way interaction of database and feature representation factors. These imply that active learning has limited success for hard databases if used with a weak feature representation.

11 and compare them for IDM and Zernike Moments). At this point, random selection might perform as well as active learning if the discrimination across classes is trivial with strong features. Hence, as the database gets more challenging, active learning will yield higher gains. This is demonstrated by smaller deficiency values (better performance) obtained on the NicIcon database compared to the COAD database when the IDM features are utilized.

A classifier requires fewer samples to become capable of making clever decisions in sample selection when the feature representation is stronger. This is depicted in Figs. 10 and 11 (compare the number of rounds required for active learning curve to get over random selection curve for IDM vs. Zernike Moments). The classifier will assume that the rough shape of the ultimate decision boundary earlier with stronger features since each sample carries more information (detail) in this case. In addition, the classifier will assign more accurate class probabilities for samples since it will check the samples in more detail (with a strong feature representation). However, as the feature representation gets weaker, it will take much more samples for active learning to beat random selection. Therefore, active learning may not yield desired performance if the feature representation is not powerful enough for the database.

We suggest being cautious in applying active learning in an unknown database and/or with a new feature representation (see Section 5.4). In particular, we recommend practitioners to take steps

Table 5
The required number of samples to reach the top accuracies that both Margin-based selection and random selection can reach. Active learning requires labeling very few samples in order to reach the top accuracies compared to random selection.

			Top (mean) accuracy(%)	Number of samples required	
				Margin-based selection[a]	Random selection
COAD	**Small Seed Set**	**Zernike**	86	270	410
		IDM	99	150	250
	Large Seed Set	**Zernike**	86	240	350
		IDM	99	130	200
NicIcon	**Small Seed Set**	**Zernike**	75	834	1120
		IDM	99	234	1100
	Large Seed Set	**Zernike**	75	846	1150
		IDM	99	246	1020

[a] The results are for Margin based selection combined with the Default batch selection strategy.

to assess the intrinsic difficulty of the database and the representative power of their feature representation and proceed accordingly. The representative power of a new feature representation can be assessed easily by trying it on major databases where the performance of the existing feature representations is well documented. Similarly, the intrinsic difficulty of a database can be assessed by training classifiers on a small subset of it (which can then be used as a seed set) using well known feature representations. Prior knowledge in the form of the intrinsic difficulty of the database and the representational power of the features can then be used to guide how active learning should be used, and whether it should be used at all as described in Section 5.4.

It is worth noting that, although we focused on two specific types of prior knowledge, we believe further research analyzing other database characteristics (style variation, noise, database size etc.) as factors will allow us to better utilize the prior knowledge on active learning.

6.3. A note on the informativeness measure

We demonstrated (in Section 5.2) that we can couple up Margin-based informativeness with Default batch selection strategy (rather than Global-FV or Global-PE) if we have scarce resources (e.g. time and computation power). However, we warn readers to be cautious as the batch size increases. Since chances of having samples with mutual information tends to increase as the number of samples in the batch increases, switching to Global-FV strategy would be safer than using Default batch selection strategy.

In conclusion, our factor analysis provides better understanding on the factors affecting active learning performance. Our results and discussion of these factors will act as a useful guide for active learning practitioners in order to make effective choices when preparing active learning setups. We hope our results will encourage future researchers to apply our methodology of factor analysis on active learning performance. As we explore more factors and understand their effects, the utility of active learning will improve.

7. Related work

To our knowledge, our work is the first extensive empirical analysis of active learning methods for sketch recognition. We presented detailed analysis on factors such as informativeness measure, batch selection strategy, feature representation, seed size and database. We also provided hints on how prior knowledge on the domain can be utilized for proper use of active learning.

There are several surveys presenting many basic and task specific active learning methods that have been proposed in the literature. However, there is a lack of extensive performance analysis. Olsson presents basic informativeness measures and approaches to active learning in a literature survey [21]. This survey also highlights concerns on active learning such as data access, re-use of annotated data, cost sensitive design and performance monitoring. In another literature survey, Settles presents query strategy frameworks and practical considerations including batch selection, noisy oracles and variable labeling costs [18]. Although both literature surveys provide detailed knowledge on many active learning methods and point out main concerns in active learning, there is no empirical study analyzing these methods extensively. We constructed a set of carefully designed experiments to investigate the performance of well established active learning methods and their combinations under different settings of various factors. In this respect, our study fills in an important niche in the active learning literature.

Although there is ample work reporting new active learning methods and their empirical performance [16,17,19,22,25], these studies simply tend to report accuracy curves, precision-recall curves or F-measure curves. They omit a rigorous statistical analysis of the results, and do not discuss how various factors interact under a variety of conditions, which we do here. In this sense, our empirical analysis methodology serves as an example to follow. Our work also creates awareness on the importance of carrying out multi-factor analysis in order to get a comprehensive assessment of various methods.

There are also lines of work that explore the effectiveness of specific active learning techniques, but they do not investigate how they interact with other factors. Settles and Craven analyze various active selection strategies with their adaptations for sequence labeling task [19]. 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 [15]. Rather than analyzing various basic methods in the literature, Markowitz analyzes variations of uncertainty sampling for large corpus labeling with boosted naive Bayesian style classifier [33]. Unlike these studies that focus on rather specific methods, we present an analysis of well established methods, but in combination with batch selection strategies. In addition, we analyze effects of factors such as feature representation, database and seed size.

There are several custom interfaces for labeling sketch data in the literature [9–12] which aim to reduce annotation effort through user friendly interfaces. Although they reduce the labeling effort required per sample, they do not consider reducing the number of samples to be labeled. We show that active learning can reduce the number of samples required to be labeled for sketch recognition. Our analysis on factors effecting active learning performance presents a valuable guide on effective use of active learning for sketch recognition. Therefore, active learning can be integrated to the available custom interfaces and further increase their effectiveness.

In other work, Plimmer et al. present a method for automated labeling of ink stroke data [13] which pre-labels data by a classifier trained with some labeled data at the beginning. Since active learning classifies the unlabeled data at each round of the process, it inherits all the benefits that automated labeling has. In contrast to automated labeling, active learning allows the classifier to

improve at each round. Hence, misclassifications will decrease over time to further reduce human effort. In this respect, the active learning framework that we have presented can be used to complement the automatic labelling interface to reap the benefits of a custom interface and a powerful mechanism for prioritizing the labelling effort.

8. Future work

In this paper, we present an empirical analysis of active learning methods and a detailed investigation that uses factor analysis. The insights gained through our analysis serve as a useful guide for users of active learning. Unfortunately, investigating a large number of factors requires huge amounts of resources in terms of time and computation power. Hence, we had to limit our analysis to five factors. Investigation of additional factors such as batch size, number of committee members, and parameter sensitivity of algorithms is likely to yield further insights into practical factors related to the use of active learning.

One unique aspect of our work is the discussion of factor interactions with respect to characteristic properties of the databases and feature representations (e.g. the NicIcon database is more challenging database, and IDM is a more powerful feature representation compared to Zernike Moments). We believe there is value in investigating factor interactions with respect to other properties as well. For instance, Sun and Hardoon apply active learning for databases which contain sparse data for some classes, but they do not investigate the effect of sparseness on active learning performance [34]. Sparseness of data, size of database and the number of classes in database are candidate properties to consider in factor analysis for future researches.

Our study considers active learning strategies targeting only effective sample selection. Future researchers might take into account the cost of labeling through various strategies and/or user interfaces. It might be interesting to apply factor based analysis in such cost-sensitive investigations of active learning. This may shed light on the interplay between sample selection strategies, batch size and labeling schemes. For example, if it is possible to devise a faster and easier labeling strategy/interface, then it might be worth considering sample selection algorithms that work better with large batch sizes.

9. Summary

In this paper, we explored the main effects and deeper interactions of various factors that impact the performance of active learning in the context of a computer graphics application that uses machine learning. In the course of doing so, we constructed an exemplary experimental design and laid out a comprehensive setup for statistical analysis of our experimental results. We investigated the performance of active learning for combinations of (a large variety of) basic informativeness measures, batch selection strategies, and the effects of factors such as feature representation, database and seed size.

As reported in the literature, active learning does not always yield the expected performance gains, and there is a need to explore the behavior of active learning for various problems. Our work is the first effort in this direction for the sketch recognition problem. We showed that for our domain of interest, Margin-based selection has superior performance. Its performance is consistent across databases, and does not require an elaborate batch selection strategy. Among batch selection strategies, we showed that Global-FV is a desirable strategy to couple with informativeness measures.

Our investigation of additional factors (i.e. seed size, database and feature representation) yielded useful insights on the selection of seed set size and utilization of prior knowledge. We demonstrated that for sketch recognition, employing larger seed sets (as resources allow) yields better active learning performance for the single classifier approach. We also showed that the use of active learning brings definitive advantages in challenging databases when accompanied with powerful feature representations.

We believe that our experiments, as well as the results that we have reported, will raise awareness on the importance of factors that impact the utility of active learning among users of sketch recognition, as well as designers of other computer graphics applications that rely on user annotation for machine learning. In this respect, in addition to the specific results reported in our experiments, our work will also serve as a valuable practical guideline that future users of active learning can follow.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.cag.2015.07.023.

References

[1] Lee W, de Silva R, Peterson EJ, Calfee RC, Stahovich TF. Newton's pen: a pen-based tutoring system for statics. Comput Graph 2008;32(5):511–24.
[2] Taele P, Hammond T. Lamps: a sketch recognition-based teaching tool for mandarin phonetic symbols i. J Vis Lang Comput 2010;21(2):109–20.
[3] Shesh A, Chen B. Smartpaper: an interactive and user friendly sketching system. Comput Graph Forum 2004;23(3):301–10.
[4] Juchmes R, Leclercq P, Azar S. A freehand-sketch environment for architectural design supported by a multi-agent system. Comput Graph 2005;29(6):905–15.
[5] Hammond T, Davis R. Ladder, a sketching language for user interface developers. Comput Graph 2005;29(4):518–32.
[6] Hou S, Ramani K. Classifier combination for sketch-based 3d part retrieval. Comput Graph 2007;31(4):598–609.
[7] Alvarado C, Davis R. Resolving ambiguities to create a natural computer-based sketching environment. In: Proceedings of the 17th international joint conference on artificial intelligence. Seattle, Washington, USA: IJCAI; 2001. p. 1365–74.
[8] Fu L, Kara LB. Neural network-based symbol recognition using a few labeled samples. Comput Graph 2011;35(5):955–66.
[9] Kaster BL, Jacobson ER, Hammond TA. Sssousa: automatically generating secure and searchable data collection studies. In: International workshop on visual languages and computing. Redwood City, CA, USA: VLC; 2009. p. 365–8.
[10] Oltmans M, Alvarado C, Davis R. Etcha sketches: lessons learned from collecting sketch data. In: Making pen-based interaction intelligent and natural. Menlo Park, CA, USA: AAAI; 2004. p. 134–40.
[11] Blagojevic R, Plimmer B, Grundy J, Wang Y. A data collection tool for sketched diagrams. In: Proceedings of the 5th eurographics conference on sketch-based interfaces and modeling. Annecy, France: Eurographics Association; 2008. p. 73–80.
[12] Wolin A, Smith D, Alvarado C. A pen-based tool for efficient labeling of 2d sketches. In: Proceedings of the 4th eurographics workshop on sketch-based interfaces and modeling. New York, NY, USA: ACM; 2007. p. 67–74.
[13] Zhen J, Blagojevic R, Plimmer B. Automated labeling of ink stroke data. In: Proceedings of the international symposium on sketch-based interfaces and modeling. Aire-la-Ville, Switzerland: Eurographics Association; 2012. p. 67–75.
[14] Settles B. From theories to queries: active learning in practice. In: Workshop on active learning and experimental design. Sardinia, Italy: JMLR; 2011. p. 1–18.
[15] Schein AI, Ungar LH. Active learning for logistic regression: an evaluation. Mach Learn 2007;68(3):235–65.
[16] Gasperin C. Active learning for anaphora resolution. In: Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. p. 1–8.
[17] Guo Y, Schuurmans D. Discriminative batch mode active learning. In: Platt J, Koller D, Singer Y, Roweis S, editors. Advances in neural information processing systems, vol. 20. Cambridge, MA: MIT Press; 2008. p. 593–600.
[18] Settles B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison; 2010.
[19] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2008. p. 1070–9.

[20] Lewis DD, Gale WA. A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: Springer-Verlag New York, Inc.; 1994. p. 3–12.
[21] Olsson F. A literature survey of active machine learning in the context of natural language processing. Technical Report T2009-06, Swedish Institute of Computer Science; 2009.
[22] McCallum A, Nigam K. Employing EM and pool-based active learning for text classification. In: Proceedings of the 15th international conference on machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. p. 350–8.
[23] Melville P, Yang SM, Saar-Tsechansky M, Mooney R. Active learning for probability estimation using jensen-shannon divergence. In: Proceedings of the 16th european conference on machine learning. Berlin, Heidelberg: Springer-Verlag; 2005. p. 268–79.
[24] Engelson SP, Dagan I. Minimizing manual annotation cost in supervised training from corpora. In: Proceedings of the 34th annual meeting of the association for computational linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 1996. p. 319–26.
[25] Shen D, Zhang J, Su J, Zhou G, Tan CL. Multi-criteria-based active learning for named entity recognition. In: Proceedings of the 42nd annual meeting on association for computational linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. p. 589–96.
[26] Brinker K. Incorporating diversity in active learning with support vector machines. In: Fawcett T, Mishra N, editors. ICML. Washington, DC, USA: AAAI Press; 2003. p. 59–66.
[27] Arandjelović R, Sezgin TM. Sketch recognition by fusion of temporal and image-based features. Pattern Recognit 2011;44:1225–34.
[28] Niels R, Willems D, Vuurpijl L. The NicIcon database of handwritten icons. In: 11th International conference on the frontiers of handwriting recognition (ICFHR 2008). Montreal, Canada; 2008. p. 296–301.
[29] Tumen RS, Acer ME, Sezgin TM. Feature extraction and classifier combination for image-based sketch recognition. In: Proceedings of the seventh sketch-based interfaces and modeling symposium. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association; 2010. p. 63–70.
[30] Khotanzad A, Hong Y. Invariant image recognition by zernike moments. IEEE Trans Pattern Anal Mach Intell 1990;12(5):489–97.
[31] Ouyang TY, Davis R. A visual approach to sketched symbol recognition. In: Proceedings of the 21st international joint conference on artifical intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2009. p. 1463–8.
[32] Baram Y, El-Yaniv R, Luz K. Online choice of active learning algorithms. J Mach Learn Res 2004;5:255–91.
[33] Markowitz TJ. An empirical evaluation of active learning and selective sampling variations supporting large corpus labeling, [PhD thesis], Pace University; 2011.
[34] Sun S, Hardoon DR. Active learning with extremely sparse labeled examples. Neurocomputing 2010;73(16–18):2980–8.