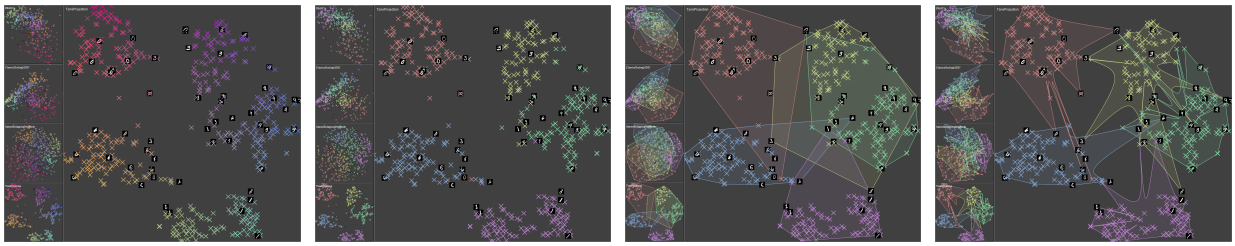
Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study

Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair



(a) 2D Colormap (b) Class Coloring (c) Convex Hulls (d) Butterfly Plot

Fig. 1: Evaluation of four visualization techniques (a)-(d) that support the visual-interactive labeling process. Our study reveals that Class Coloring (b) and Convex Hull (c) are the most useful techniques. Both capture characteristics of the input data and the classification model in an intuitive way. Our study shows that they can compete with and even outperform active learning strategies.

Abstract—Labeling data instances is an important task in machine learning and visual analytics. Both fields provide a broad set of labeling strategies, whereby machine learning (and in particular active learning) follows a rather model-centered approach and visual analytics employs rather user-centered approaches (visual-interactive labeling). Both approaches have individual strengths and weaknesses. In this work, we conduct an experiment with three parts to assess and compare the performance of these different labeling strategies. In our study, we (1) identify different visual labeling strategies for user-centered labeling, (2) investigate strengths and weaknesses of labeling strategies for different labeling tasks and task complexities, and (3) shed light on the effect of using different visual encodings to guide the visual-interactive labeling process. We further compare labeling of single versus multiple instances at a time, and quantify the impact on efficiency. We systematically compare the performance of visual interactive labeling with that of active learning. Our main findings are that visual-interactive labeling can outperform active learning, given the condition that dimension reduction separates well the class distributions. Moreover, using dimension reduction in combination with additional visual encodings that expose the internal state of the learning model turns out to improve the performance of visual-interactive labeling.

Index Terms—Labeling, Visual-Interactive Labeling, Information Visualization, Visual Analytics, Active Learning, Machine Learning, Classification, Evaluation, Experiment, Dimensionality Reduction

1 Introduction

Labeling follows the principle of attaching information to some object. In data-centered disciplines labeling is often associated with querying knowledge of users about data objects. As such, the labeling process represents an essential prerequisite for algorithmic support in data mining, machine learning, and visual analytics. Two goals of almost any labeling process are being accurate and fast, i.e., effective and efficient.

In the machine learning community, labeling traditionally represents the basis for the creation of large ground truth data sets. Ground truth is necessary to enable autonomous supervised learning. The most recent and powerful supervised machine learning approaches, such as deep neural networks require large amounts of such labeled data to learn successfully. The generation of such datasets is, however, expensive and often requires extensive efforts from the users (e.g. crowdsourcing [28]). Active learning (AL) is one promising approach to reduce the labeling effort. The basic principle of AL is to query an oracle (the user) for labels about individual objects (instances) in the dataset. Thereby the active learner selects those candidates from which the classifier is expected to benefit most. Various AL strategies have been proposed and shown to improve over random sample selection [56].

One of the intrinsic characteristics of AL strategies is the *model-driven* way to identify meaningful instances for labeling. A drawback of this principle is that users, with their ability to identify patterns very fast, have no influence on the candidate selection. Considering the efficiency, strategies asking users for a single or multiple labels in an iterative manner does not scale well for large data sets. Finally, a particular challenge for model-driven strategies is the cold start (bootstrap) problem, i.e., starting the learning with no labeled instances at all [37]. The question arises whether or not classical AL approaches can benefit from visual interfaces that take the user into the loop, not only for labeling but particularly for *selecting* meaningful candidates [55, 57].

In the visual analytics community, interfaces for visual-interactive labeling (VIL) become increasingly popular since they enable users to express their information need. Additionally, visual analytics models can support the users' knowledge generation process by exploiting such label information. Approaches differ in the type of acquired labeling information ranging from interestingness scores, rules, similarity relations to assignments of class labels. One common ground of many approaches is the *user-driven* selection of instances provided with

- *Jürgen Bernard is with Technische Universität Darmstadt, Darmstadt, Germany. E-mail: juergen.bernard@gris.tu-darmstadt.de.*
- *Marco Hutter is with Technische Universität Darmstadt, Darmstadt, Germany. E-mail: marco.hutter@gris.tu-darmstadt.de.*
- *Matthias Zeppelzauer is with St. Pölten University of Applied Sciences, St. Pölten, Austria. E-mail: matthias.zeppelzauer@fhstp.ac.at.*
- *Dieter Fellner is with Fraunhofer IGD, Darmstadt, Germany. E-mail: dieter.fellner@gris.tu-darmstadt.de.*
- *Michael Sedlmair is with University of Vienna, Vienna, Austria. E-mail: michael.sedlmair@univie.ac.at.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

visual interfaces for the exploration and identification of interesting instances. The motivation of users for the selection of particular instances can be diverse. It may be task-dependent but also the context, application field, and more intrinsic backgrounds of users may motivate the instance selection. These different motivations may lead to biased selections of instances that result in suboptimally trained models. Following the principle of AL strategies, guiding users in the candidate identification may be beneficial to mitigate subjective or suboptimal instance selection.

Our research is motivated by observing the different strengths and weaknesses of the respective principles and the idea of a substantiated combination of mutual strengths in future approaches. Recent developments in machine learning and visual analytics indicate that the two fields are getting closer, see for example a recent survey that exposes the considerable overlap [43]. Furthermore, there have been first initiatives to combine active learning and visual-interactive label selection [22, 23, 55, 57]. We share the vision that the strengths of both principles can be seamlessly combined in visual-interactive labeling systems, raising the effectiveness and efficiency to new levels.

From our literature research, we conclude that the strengths and weaknesses of both principles have hardly been assessed in direct comparison. There is a lack in formalizations of AL and VIL strategies and more generally of labeling approaches in machine learning and visual analytics. Thus, reliable decisions on whether or not to choose one or the other approach are hardly feasible. A core driving question is whether or not visual interfaces can improve labeling tasks, or whether AL performs so well that they render visual interfaces redundant. In case that VIL is really helpful, other interesting questions arise: Do users have particular strategies for the identification of labeling candidates? How good are these strategies with respect to their performance? Under which circumstances does VIL help, and when not? How does VIL perform for differently complex tasks? And how much and what information about data and machine learning models should be represented in visual interfaces? These questions have not been answered yet and require a closer investigation.

To answer these questions and to provide a direct comparison of VIL and AL labeling strategies, we perform an experimental study. As a basis for the study, we develop a flexible evaluation toolkit that integrates 16 different established AL strategies, five classifiers and four visualization techniques (Sec. 3). Using this toolkit, we conducted an empirical study with 16 expert participants. Our study sheds light into (i) how VIL and AL techniques compare to each other, (ii) how the complexity of the labeling situation impacts them, and (iii) the differences between single- and multi-instance labeling approaches (Sec. 5.1-5.2). We also (iv) characterize a set of labeling strategies that we found our participants applying in VIL conditions (Sec. 5.3). With these findings, we discuss lessons learned, insights gained, and potential future work (Sec. 6).

Our investigation shows that VIL achieves similar performance to AL and in some settings even outperforms AL. It further points out new connecting points where VIL and AL may benefit from each other. The presented investigation represents an important step towards a unified labeling process that combines the individual strengths of user-centered and model-centered strategies.

We present related work in the next section, followed by our baseline approaches in Section 3, used for our experimental study. In Section 4, we introduce the experiment design, and present the experiment results in Section 5. We discuss follow-up insights in Section 6, and conclude with a discussion on limitations and future work.

2 Related Work

Related work for our study comes from different domains. Thus we subdivide the presentation of related work into three sections: related work on active learning (Section 2.1), visual interactive labeling (Section 2.2) and previous studies on visual interactive labeling (Section 2.3)

2.1 Active Learning

Active learning (AL) is a special type of semi-supervised machine learning which takes the user into the loop to query label information to improve the training performance of a classifier. AL techniques ask (query) an oracle (the user) for specific instances instead of, e.g., querying random instances. AL is especially useful in cases where large portions of the data are unlabeled, or where manual labeling is expensive. Thereby, the major goal of AL is to achieve high accuracy with a minimum of manual labeling effort. The core component of AL is the candidate selection strategy which aims at identifying those instances which would contribute most to the learning progress of the model. The different classes of AL strategies are described in several surveys in detail [39, 56, 65, 68]. We partition AL strategies into four major classes: (i) uncertainty sampling, (ii) error reduction schemes, (iii) relevance-based selection, and (iv) purely data-centered strategies.

Uncertainty sampling aims at finding the instances that the learner is most uncertain or unsure about. A widely used strategy is to search for those samples near the decision boundary of margin-based classifiers [72] also referred to as *large-margin based AL* [65]. Other strategies measure the uncertainty of a committee of classifiers. In *Query by Committee (QBC)* [60], each classifier of the ensemble is asked for labelings. Instances are considered interesting when the committee disagrees with respect to their labeling [36].

Error reduction schemes focus on the selection of those instances which may change the underlying classification model most. Techniques focus either on the impact on the training error (expected model change) [59] or on the reduction of the generalization error (risk reduction [42] and variance reduction [24]).

The third group of AL strategies focuses on *relevance* [67]. Based on a relevance criterion, those instances are selected which have the highest probability to be relevant for a certain class. This strategy fosters the identification of positive examples for a class. This is particularly useful in systems that aim at ranking search results [68].

Finally, one of approaches is purely *data-driven* and independent of the learning model. Examples for such data-driven strategies are density- and diversity-based instance selection. The diversity criterion fosters the selection of dissimilar instances for labeling to increase the information gain for the learner [17]. In density-based selection, the query candidates are selected from dense areas of the feature space because those instances are considered as most representative [72]. Density-based selection of candidates is a promising strategy for initiating an AL process in the case when no labels are available at all (cold start problem).

In this work, we employ a heterogeneous set of 16 AL strategies to obtain a representative baseline for AL (see Section 3.3).

2.2 Visual-Interactive Labeling and Classification

Labeling is a frequently supported task in visual analytics. Depending on the given task and approach, different types of labels may be employed. A widely used label type are categorical labels which can be either binary or multi-valued. Binary labels enable simple user feedback, such as "yes/no" decisions or "relevant/not relevant" assessments. Multi-valued categorial labels enable for example the tagging of different classes of objects. Users can, e.g., label relevant textual documents [22], interesting time series patterns [47], or occurrences of objects in video streams [23]. Another important label type are continuous labels, often used to assign more fine grained interestingness or relevance scores. Example applications include relevance feedback [45, 54], candidate assessment and evaluation [71], patient well-being scores [4], or distinguishing between relevant and irrelevant views [2]. Aside from providing labels explicitly, another type of user feedback is to directly provide weights (of features or data attributes), e.g., to build, validate, or improve algorithmic models [38, 40, 70]. Another type of label are similarity relations between instances, explicitly assigned by users which are used, e.g., to learn distance functions to support the visual-interactive re-allocation of instances [5, 10, 35].

Labeling is an upstream task for (visual-interactive) classification approaches. Some approaches directly combine AL-based with VIL-based instance identification and labeling [22, 23, 55, 57]. Seifert and Granitzer [55] elaborate on user-picking strategies similarly as we do in our experiment. In contrast, the baseline interface does not aim for a similarity-preserving representation of instances. Heimerl at al. [22] and Höferlin et al. [23] present visual analytics systems

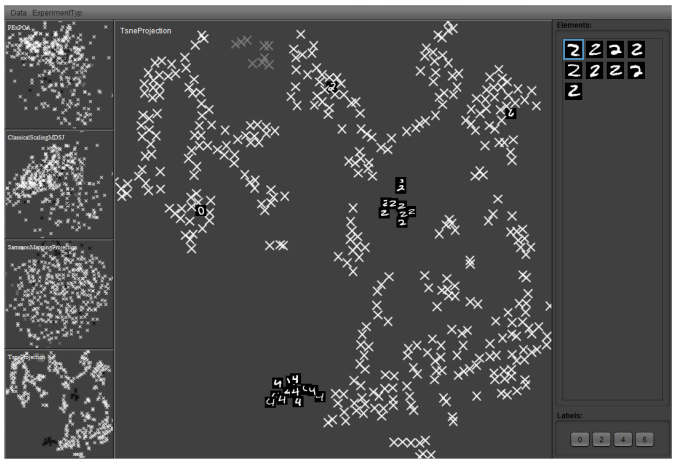


Fig. 2: Visual-interactive interface for labeling instances. Four dimensionality reduction techniques provide different perspectives on the data set (default: t-SNE). In the center instances can be selected for labeling. At the right, users can refine selections and label multiple instances at once (in $\mathbf{TR}_3$). Four different VIL-support techniques can be included to ease the visual-interactive labeling process (see Figure 1).

combining multiple views including VIL support, model visualization, and instance labeling. We build upon these approaches for the implementation of our study results. Unsupervised techniques can be employed to ease the task of user-based labeling. Several approaches provide visualizations of cluster results in 2D, in combination with interaction tools like a 'lasso' for multi-instance selection [12, 21, 38]. From clustering, we take up the idea to support users in labeling instances in dense areas that are most representative. In this context, we will also examine whether or not visual cluster structures and user expectations [38] are beneficial for the labeling process.

2.3 Previous Studies on Visual Interactive Labeling

The number of experiments and studies regarding the performance of VIL is scarce. At a glance, our experiment builds upon insights gained from studies on the identification of human labeling strategies, the comparison of labeling-support techniques, multi-instance labeling, as well as measures of class separability. Möhrmann et al. measured the applicability of SOM-based data clustering and visualization as a means to support the generation of ground truth for image data [38]. Similar to one of our experiment trials, the authors assess the increase in efficiency when labeling multiple instances at once. However, the authors report accuracy values that remained constant with a slight tendency to deterioration. The experiment conducted by Settles measures the annotation time of users versus accuracy [57]. We build upon the ideas to raise baseline random performance measures as well as an upper limit of performance to provide upper and lower bounds. In contrast to our experiment, the comparison is between learning from instances, with and without additionally learning from features. In their studies, Seifert and Granitzer [55] simulated user-picking strategies for instance selection, allowing the automation of user-based selection in a laboratory study. The authors presented a VIL-support technique based on radially ordered axes of a classifier's a-posteriori output probabilities and claimed that their technique outperforms uncertainty based sampling (AL) [56].

In contrast to previous studies, we further focus on the comparison of model-based (AL) versus human-based (VIL) label selection strategies. In particular, we are interested in observing what drives users to select particular instances in a given labeling interface. One way for humans to enhance the labeling process is the ability to identify patterns such as dense areas of instances and class distributions. Our experiment builds upon the results of a study on visual cluster and class separability [53]. With the evaluation of techniques supporting VIL, we further investigate the effect of well-separable class distributions on effective and efficient labeling. One inspiring side-aspect is entailed in an experiment comparing the results of cluster validity measures with user evaluations (experts and non-experts) [33]. Building on the basic assumption that model-based and user-based strategies of candidate identification differ, we will adopt the idea to further observe user preferences in candidate selection. Finally, Lewis et al. [34] conducted an experiment on whether humans are consistent in rating the quality of results of dimensionality reduction algorithms. Similarly, a motivating aspect for our user experiment is to investigate the consistency of user-based labeling strategies.

3 Approaches

The major goal of this work is to compare AL and VIL strategies for data labeling tasks. For this purpose we have developed a toolkit that allows for simulating AL experiments as well as performing visual interactive labeling of data by users (identification, selection, and labeling of instances). Section 3.1 provides more details. We integrate a number of AL techniques and classifiers into the toolkit which we summarize in Sections 3.3 and 3.2. The visualization techniques that we propose to support VIL strategies are described in Section 3.4. We refer to them as *VIL-support techniques* in the following.

3.1 Visual-Interactive Labeling Toolkit

We present a visual-interactive toolkit that supports the visualization of high-dimensional datasets, the integration and (automated) evaluation of AL strategies, and the enrichment with VIL-support techniques to ease the labeling process. The interaction loop of the labeling process builds upon the visual-interactive labeling process presented by Bernard et al. [7] which also contains a graphical representation of the process. The design of the visual-interactive labeling interface fulfills the following three primary requirements: First, the toolkit provides a visual representation of the entire data set in 2D in a structure-preserving way. Second, the interaction design facilitates the selection and labeling of single and/or multiple instances. A lasso-tool allows the selection of multiple instances, i.e., a range selection in the 2D data representation. Third, to enable the objective comparisons between AL and VIL, labeling interactions by users trigger the same mechanisms for model building and performance testing as automatically executed AL strategies.

The visual interface of the toolkit is presented in Figure 2. Inspired by experiments on dimensionality reduction, scatterplots, and measures of class separation [52, 53, 63], our toolkit uses dimensionality reduction techniques to map high-dimensional data into 2D. A total of four techniques (PCA [27], non-metric MDS [30], Sammons Mapping [46], and t-SNE [66]) are used in a small-multiples setting to provide different perspectives on the data. This mitigates weaknesses of individual techniques. An overview of dimensionality reduction techniques for visualization is provided by Sacha et al. [44], parameter values used for the four techniques are described in the supplemental material.

The three primary views of the labeling interface are as follows. In the left view, users can select one of the four dimensionality-reduced data representations (default: t-SNE). The selected mapping is subsequently presented in the center view. At the beginning, all data instances are represented with small crosses in the center view, indicating that they are unlabeled. Once users label individual instances, the instances are depicted with small visual representations, in our case thumbnail images. The right view allows the refinement of selected subsets and multi-instance labeling, triggered by the label buttons at the bottom. The labeling information is then fed back to the underlying machine learning models which are re-trained on the enriched training set. The interaction loop is closed as soon as the results of the learning model are finished. The new classifier predictions are propagated to the center view which is updated with the new results [7]. Section 3.4 provides details about visualization techniques used to represent the classification output. Note that the same process is performed for AL strategies.

3.2 Classifiers and Classification Accuracy

We integrate five different classifiers into the toolkit (Support Vector Machine (SVM) [13], Random Forest (RF) [9], Naive Bayes [18], Multilayer Perceptron (MLP) [25] and Simple Logistic [31]). The classifiers are used for testing the performance of labeled sets of instances in combination with the learned models in our study. With the use of five different classifiers, we achieve robustness in the assessment of

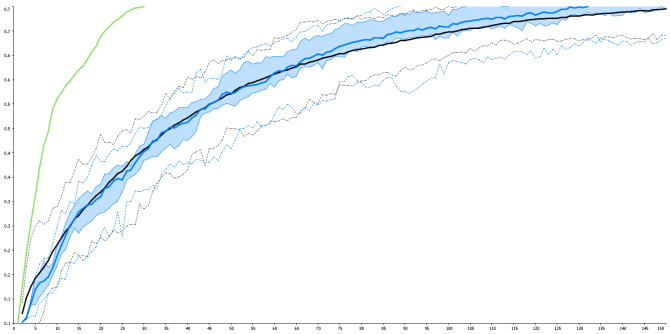


Fig. 3: Accuracy of performance baseline strategies developing over 150 labeled instances (x-axis). The performance of 50 baseline random sampling trials (black) is compared with 16 variants of AL strategies (blue). Dashed lines represent the minimum and maximum performance, filled areas depict 25% and 75% quartiles. In the result, we identify a frequently observed pattern: AL strategies start poor (cold start problem), but outperform the random baseline in later phases. The upper limit of performance (green) performs exceptionally well.

labeling performance. The computation of classification accuracy is always performed on a separate instance set for testing [19].

3.3 Active Learning Strategies

In active learning (AL) an algorithmic model proactively asks the user for feedback (labels) about selected samples [56]. Basically, the algorithmic selection of unlabeled instances is based on the result of an included classification model in combination with a quality criterion (see Section 2.1 and the supplemental material for details about the operating principles of AL). We integrate 16 supervised AL strategies into the evaluation toolkit that build upon eight different AL techniques. Techniques include Smallest Margin [48], Entropy-Based Sampling [58], Least Significant Confidence [14], Simpson Diversity [62], Probability Distance, Vote Comparison, Vote Entropy [16, 61], and Average Kullback Leibler [37]. The first four techniques are combined with the three classifiers (SVM, MLP, RF) each, yielding 12 different AL strategies. The latter four techniques are query-by-committee (QBC) approaches that use all three classifiers simultaneously and vote over their individual decisions. This adds up to a total of 16 AL strategies.

3.4 VIL-Support Techniques

Related work in visual classification approaches reveals a series of techniques that have the potential to support VIL. In this work, we define a *VIL-support technique* as a visualization that assists the user in the selection of candidates for labeling, i.e., VIL-support techniques may facilitate *VIL strategies*. We present four VIL-support techniques that we assume to be particularly interesting and beneficial for interactive labeling. Enlarged example figures of all four VIL-support techniques are provided in the supplemental materials. The labeling toolkit presented in Figure 2 serves as the baseline for all techniques. We apply voting between the five classifiers included in the system to condense the information of multiple classification results to a single class prediction for every instance. Accordingly, this condensed information about the current state of the learning models is exploited with the VIL-support techniques.

2D Colormap Every data element is colored with respect to its 2D location in the projection shown in the center view. The colors are linked to the small multiples at the left to support the lookup of instances in other views and the comparison of different perspectives on the data. In this way, users can expose mapping distortions and thus make informed decisions when selecting data elements. We use a 2D colormap to represent position information with continuous and similarity-preserving colors [6] (Figure 1a). Compared to the remaining techniques, this VIL-support technique does not require any information about the underlying classification model.

Class Coloring Each class is assigned a separate color [52, 53, 63]. Coloring classes or clusters in scatterplots is a frequently applied approach [10, 22, 23, 41, 51, 55, 64]. Here we evaluate if this technique is also beneficial to support VIL (Figure 1b).

Convex Hull The convex hull is a prominent technique for the visualization of class distributions and boundaries [41, 49, 51, 55, 64]. We employ convex hulls to visualize the boundaries of the classes (Figure 1c) and investigate its suitability for VIL.

Butterfly Plot The butterfly plot technique [50] is an interesting refinement of convex hulls that additionally provides information about the center of gravity of class distributions (Figure 1d). While the butterfly plot tends to produce more complex shapes it better highlights outliers than convex hulls.

4 Experimental Design

We conducted an experiment with three main distinctive parts: (**PART**$_{1\text{-}3}$). Each part focused on a specific set of questions. The major goal of our experiment was to examine the potential of VIL and how it compares to AL. We first describe the general setup of the experiment, before we provide details for each part (variables, setup, tasks, etc.).

4.1 Research Questions

We formulated six research questions for our experiment:

- **RQ**$_1$ – Is VIL, facilitated with VIL-support techniques, able to compete with state-of-the-art AL strategies?
- **RQ**$_2$ – Is VIL effective even in complex labeling settings?
- **RQ**$_3$ – Do VIL-support techniques perform differently?
- **RQ**$_4$ – Can VIL facilitate the concept of labeling multiple instances at once, to make the process more efficient?
- **RQ**$_5$ – Do users develop strategies for the selection of meaningful instances in VIL?
- **RQ**$_6$ – How do these potential strategies relate to VIL-support techniques and AL strategies?

4.2 Baseline AL Strategies

To obtain a representative and robust baseline a broad range of existing state-of-the-art approaches is required. These approaches can be run for comparison automatically and do not need to be tested by users directly, so we are not restricted by the participants' time here. We thus selected 16 AL strategies as baseline conditions (cf. Section 3.3).

4.3 Data Set

To keep the study complexity manageable, we use a single, easy-to-understand reference data set in our study. After reviewing a number of data sets the decision was made for the MNIST data set representing classified handwritten digits [32]. The database contains 60,000 instances for training and 10,000 instances for testing from 10 distinct classes (digits "0" to "9"). Each raw digit is represented by a 28x28 grayscale image yielding a 784 dimensional vector in the original space. The grayscale values represent the luminance information of the digits, while black color encodes the background (see, e.g., Figure 2 in the top-right corner). To reduce the dimensionality for faster classification, we extract a descriptor based on 11 horizontal, 11 vertical, and 20 diagonal slices carved out from the original grid. A detailed description of the feature extraction is provided in the supplemental materials. The final feature vector is applied as input for training and testing classifiers, executing AL strategies, and applying dimensionality reduction.

4.4 Participants

We recruited 16 participants (2 female) in our lab. Each participant performed all three parts of the experiment. The age of the participants ranged from 26 to 58 (Median = 33.06, SD = 7.56). All participants had normal or corrected-to-normal vision. Each subject had at least a Bachelor's degree and expertise in visualization, data mining, machine learning, or combinations thereof. However, none of the participants has either worked with the particular data set in detail, or has in-depth experience in implementing classifiers.

4.5 Procedure

We prepared a workstation in a quiet lab with a color-calibrated monitor. The evaluation toolkit (cf. Section 3.1) was installed and prepared for the experiment. Figure 2 gives an impression of the toolkit. By design, all unlabeled (unknown) instances are represented with *x*, labeled instances are depicted with the image of the handwritten digit.

At the beginning, the participants were introduced to the topic and the goals of our experiment, accompanied by the possibility to ask questions. Our toolkit was introduced in a short demo session, including its interaction techniques and VIL-support techniques. In addition, the concept of (baseline) AL strategies was described, as well as the functionality of the classifiers to be trained in the course of each session.

The main part of our study consisted of three core parts then:

- $\mathbf{PART}_1$ – Users were asked to label the data under the 4 different VIL conditions described in 3.4. Our goal was to learn about how VIL techniques compare among each other, as well as to the baseline AL strategies, and how they do so in differently complex situations.
- $\mathbf{PART}_2$ – The users had to engage in a single and a multiple-instance selection tasks, so we can compare single vs. multiple instance labeling strategies for AL and VIL.
- $\mathbf{PART}_3$ – We gathered qualitative and subjective feedback from the participants.

During $\mathbf{PART}_1$ and $\mathbf{PART}_2$, the participants were asked to think aloud in the course of the labeling process, e.g., when they identify special cases, difficulties, or interesting findings. We also observed them and took notes on interesting behaviors and user strategies. Both parts used a separate within-subject design, which will be described in more detail below.

The overall time to perform the three parts was estimated with 75 minutes, depending on the extent of the interview. Participants were allowed to take breaks between the three parts. We now describe the experimental design of each part in more detail.

4.6 $\mathbf{PART}_1$: Performance Comparison VIL and AL

The first part of the experiment considered the question whether or not the four VIL-support techniques can compete with state-of-the-art AL strategies, how they compare among each other, and how they perform in three different levels of complexity ($\mathbf{RQ}_{1\text{-}3}$). To answer these questions, $\mathbf{PART}_1$ was organized as a 4×3 within subject design.

4.6.1 Independent Variables

We had two independent variables in $\mathbf{PART}_1$.

VIL. Our main variable of interest were the *4 different VIL-support techniques* as outlined in Section 3.4: *2D Colormap*, *Class Coloring*, *Convex Hulls*, and *Butterfly Plot* (cf. Figure 1).

Complexity. The second variable was *task complexity*. Complexity of the labeling task at hand is a very important factor that can strongly influence AL and VIL performance. Task complexity in itself, however, is a multifaceted concept. It is influenced by model aspects such as the number of different classes, how many data points it is operating on, the chosen model type, etc. It also depends on the input data and the nature of the labeling tasks, for instance, labeling digits might be easier than labeling objects in video streams. No single study can investigate all of these factors at once. Based on our pilot study (see suppl. materials), we thus opted for a well-defined labeling task (labeling digits), and focus on three different levels of model complexity:

1. *Easy*: 2 classes (0,1), 100 instances each class
2. *Medium*: 5 classes (0,1,2,3,4), 100 instances each class
3. *Difficult*: 10 classes (0,1,...,9), 100 instances each class

4.6.2 Task Description

We asked the participants to select data instances for labeling in a meaningful way based on their preference. Depending on the provided VIL-support technique additional information about data and/or classification result was shown that possibly supported the process of selecting instances for labeling. As a general rule, we asked the participants to exploit relevant information about patterns explored in the labeling interface, and use it for the selection (labeling) of instances. For every condition, users were informed about the set of labels existing in the data set (task complexity).

The focus of $\mathbf{PART}_1$ was on the *selection* of instances rather than the actual process of assigning a label to the selected instance. We thus setup our toolkit in a way that participants did not actually need to label the digits of selected instances to save time. They could simply select an instance by clicking on it. The label was then set automatically and the image was revealed to the user. The registration of the true label of for the selected instance was automatically triggered to the evaluation bench (see the interaction loop in Section 3.1).

4.6.3 Setup

The independent variables of $\mathbf{PART}_1$ lead to $4 \times 3 = 12$ different conditions. We decided for a within-subject design, so all 16 participants were asked to perform all 12 conditions. We decided not to randomize the order of VIL techniques as the conditions are building up on top of each other (the number of visual variables depicting model information was zero for *2D Colormap*, one for *Class Coloring*, and two for *Convex Hulls* and *Butterfly Plot*). The three different complexities of the labeling task were always performed in the natural order from easy to difficult. The data instances used for training and testing were randomly chosen with a constant seed to achieve both comparability and reproducibility. All other choices were based on the pilot study that we describe in Appendix A in the supplemental materials.

4.6.4 Dependent Variables

Accuracy. To assess the performance of the 12 tested conditions outlined above, a measure is needed that is expressive and easy to understand. To enable comparability, the measure should be applicable for the 12 conditions and the baseline AL strategies. Based on these requirements, we select classification *accuracy* as the sole dependent measure to compare how 'good' the different conditions are. We use the standard definition of classification accuracy, that is, the portion of correctly classified instances compared to ground truth labels [19]. To achieve robust (classifier-independent) accuracy estimates, we compute the accuracy after every label operation for all five classifiers listed in Section 3.2 and average the results. This leads to robust performance estimates.

4.7 $\mathbf{PART}_2$: Labeling Single vs. Multiple Instances

In the second part of the experiment, we turn towards the assessment of efficiency of the labeling process. To this end, we allow labeling multiple instances with the same label in a single labeling iteration. We investigate whether or not VIL can facilitate the concept of labeling multiple instances at once to make the process more efficient ($\mathbf{RQ}_4$).

4.7.1 Independent Variables

Single vs. Multi Labeling. We were interested in the question of how many labels should be set at once in the labeling interface. There are essentially two options. Setting one label to a single instance, one after another, or assigning a label to multiple instances at once.

VIL vs. AL Strategies. Assigning labels to multiple instances at once can be facilitated with AL and for VIL as well. As a result, and in contrast to $\mathbf{PART}_1$, AL now shifts from an automated baseline approach to an independent variable as the participants need to get active in these conditions as well.

4.7.2 Task Description

In contrast to $\mathbf{PART}_1$, the users' task in $\mathbf{PART}_2$ was to explicitly *assign labels* to instances. That is, users selected and labeled instances (in VIL conditions), or they labeled suggested instances (in the AL condition).

4.7.3 Setup

Altogether, the two independent variables form four conditions. We refer, to these four different conditions as:

- *AL single labeling*: AL suggests one item, labeled by the user
- *AL multi labeling*: AL suggests multiple items to be labeled by the user; the user can select a subset of the suggestions and label them with one class label, for example, all "1s".

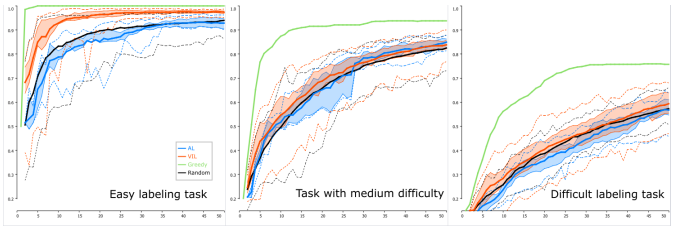


Fig. 4: The accuracy of labeling strategies depends on the complexity of the labeling task. The performance of all VIL results (orange) can at least compete with the performance of AL (blue), and RB (black). ULoP (green) substantially outperforms all remaining strategies. Dashed lines represent minimum and maximum performances, area dyed with decreased alpha is used to depict 25% and 75% quartiles.

- *VIL single labeling*: The user selects a single item and labels it.
- *VIL multi labeling*: The user can select multiple instances with a lasso and give a label to this selection after filtering false positives.

We provide dedicated interfaces for these conditions. For the VIL conditions, we use the interface described in Section 3.1. For VIL single, the user can only click select single items; for VIL multiple, he can use the lasso as described in Section 3.1. The interface was based on the *Convex Hull* design option, and is shown at the right of Figure 2.

In case of *AL multi-labeling*, we introduce a list-based interface, which allows the user to see the AL-suggested instance(s) and label them. In the AL single case, only one instance at a time is shown. In the AL multiple, multiple instances are shown and the users can select the items they think being to a certain class, and label them. We chose Smallest Margin [48] in combination with a Support Vector Machine (SVM) [13] classifier from the set of 16 AL strategies (cf. Section 3.2), as it is well-known, easy to implement, and produced consistently robust results with accuracies above average. Screenshots of all interface conditions are in the supplemental material.

We decided again for a within-subject design, so every participant was asked to perform all four conditions. The order of the conditions was from simple to difficult regarding labeling single or multiple labels at once. The order of VIL and AL was randomized. The data instances used for training and testing were again randomly chosen with a constant seed to achieve both comparability and reproducibility. The difficult labeling task was chosen (cf. Section 4.6), thus, all labels from 0 to 9 were included in the data set. For each condition, participants were asked to label as many instances as possible in 5 minutes.

4.7.4 Dependent Variables

Accuracy. The accuracy measure described in $\mathbf{PART}_1$ is again used to assess the *effectiveness* of the labeling task.

Number of Labeled Instances. In addition, we are interested in a performance measure assessing the *efficiency*. Thus, we also look at the number of instances labeled over time as a second dependent variable.

4.8 $\mathbf{PART}_3$: User Strategies and Feedback

After the two main parts of the study, the moderator conducted a summative interview, including questions about preferences, informal feedback, and subjective estimates about the usefulness of VIL-support techniques. We also handed out a short questionnaire to gather additional subjective feedback, with 5-point Likert scales regarding the subjective preference on VIL-support techniques.

The rationale of this part ($\mathbf{PART}_3$) was to answer $\mathbf{RQ}_5$ and $\mathbf{RQ}_6$, that is, whether or not users developed strategies for the selection of meaningful instances in VIL, and how these potential strategies relate to VIL-support techniques and AL strategies. Inspired by the algorithmic formalization of AL strategies for the selection of instances, we sought for formalizations of strategies performed by users when selecting instances for labeling. This was further informed by the qualitative input we got from the think-aloud protocols and our qualitative observations

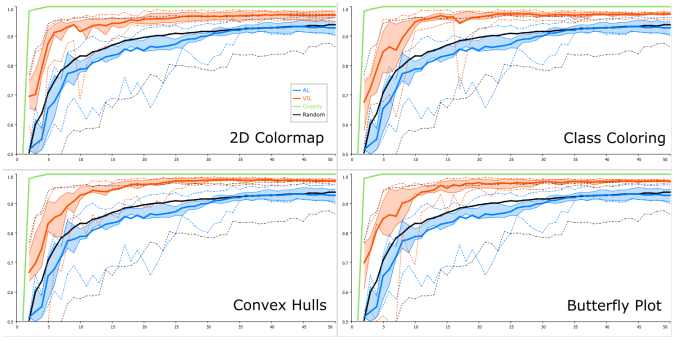


Fig. 5: Performance comparison of four VIL-support techniques (orange) with AL strategies (blue), RB (black), and ULoP (green) for an easy labeling task ($\mathbf{PART}_1$). The overall insight is that all VIL-support techniques outperform AL and RB. In many cases the accuracy was around 0.95% after only three labeled instances.

(see Section 4.5). All sessions were audio-recorded. We used a lightweight open/axial coding approach to analyze this qualitative data [11]. The analysis was done by one of the authors.

4.9 Data Analysis

To analyze our data, we mostly leverage visual representations of the performance of the different VIL and AL strategies over the course of many iterations, that is, line charts. Variations in the performances of the 16 participants (and 16 AL strategies) are visually represented as a "bundle" with an emphasis on statistical information represented over time. The colored area around a mean curve represents the interquartile range of the measured results ($[Q_{0.25} - Q_{0.75}]$), dyed with decreased alpha. Dashed line charts depict minimum and maximum performances. As a general rule, the color coding used to assess performance of curves is orange for VIL and blue for AL. We superimpose AL and VIL results to allow for an easy visual comparison (i.e., comparison of bundles).

We also wanted to contextualize our findings by providing upper and lower bounds in the experiment. We thus provide two additional pieces of information in the line charts: a *random baseline (RB)* shown in black, and an *upper limit of performance (ULoP)* shown in green. For RB we simply sample the items for labeling randomly. To achieve robustness, 50 RB trials are calculated for every evaluation. The expectation is that the remaining strategies should at least perform better than this RB. For a similar purpose, we provide ULoP, which simulates an optimal labeling strategy where always the "best" (most beneficial) item is selected in each iteration. The calculation is based on a Greedy search, simulating the accuracy of the next labeling step for all remaining candidate instances. Figure 3 shows the results of the AL strategies compared to RB, and the ULoP line chart, and illustrates our approach of visual data analysis.

We furthermore use confidence intervals (*CI*) for our analysis, following APA's up-to-date recommendation for statistical analyses [1]. In the following, we use M for the sample mean as well as CI for the confidence interval defined by $M \pm Z_{score} * SD/\sqrt{n}$ [15]. We define $Z_{score} = 1.96$, representing the commonly used $CI = 95\%$.

5 Results

We report results for the visual-interactive user experiment structured in three different parts. In Section 5.1, we take the factors of varying complexities of labeling tasks and different VIL-support techniques into account ($\mathbf{PART}_1$). Results of the comparison of single and multiple instance labeling tasks are presented in Section 5.2 ($\mathbf{PART}_2$). Finally, in Section 5.3, we report insights gained from the observation of participants, summative interviews, and informal feedback ($\mathbf{PART}_3$). Large figures of all results are provided assupplemental material.

5.1 $\mathbf{PART}_1$: Performance Comparison VIL and AL

The questions answered in $\mathbf{PART}_1$ are whether or not VIL is able to compete with state-of-the-art AL strategies in different conditions

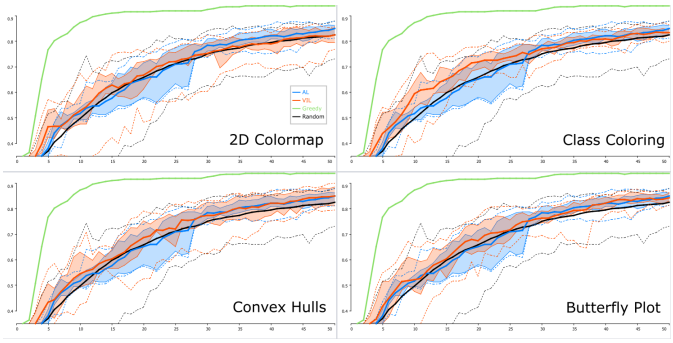


Fig. 6: Performance comparison of four VIL-support techniques (orange) with AL strategies (blue), RB (black), and ULoP (green) for a labeling task with medium difficulty ($\mathbf{PART}_1$). All VIL-support techniques have accuracies at least as high as AL and RB in earlier phases of the process. Convex hulls can compete with the AL performance for the entire observed process. ULoP (green) substantially outperforms all remaining techniques.

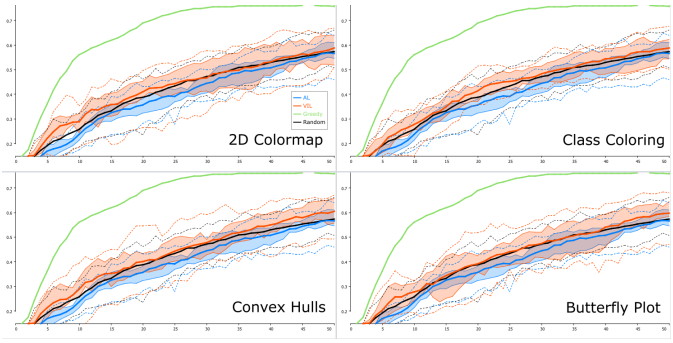


Fig. 7: Performance comparison of four VIL-support techniques (orange) with AL strategies (blue), RB (black), and ULoP (green) for a difficult labeling task ($\mathbf{PART}_2$). For difficult labeling tasks, the performance of AL and VIL is more similar, with VIL having slight advantages over AL. RB performs surprisingly well. All VIL-support techniques have a good starting performance. Convex Hulls and Butterfly Plots perform well over the entire labeling process.

($\mathbf{RQ}_{1,2,3}$). To that aim, we investigate the performance of four VIL-support techniques in combination with three different task complexities (details about $\mathbf{PART}_1$ in Section 4.6).

5.1.1 Performance for Different Task Complexities

First, we assess the dependency of the labeling performance on the complexity of the labeling settings. Three different task complexities *easy*, *medium*, and *difficult* are evaluated and compared.

Results Figure 4 shows the overall results of the performance comparison for *easy* (left), *medium*, and *difficult* (right) labeling tasks. In general, the performance of virtually any strategy increases in the course of the labeling process. Please note that the performances of the four VIL techniques are aggregated for every complexity level. The results show that from *easy* to *difficult* the *accuracy* decreases substantially for all strategies, i.e., we infer that the complexity level has an influence on the labeling performance. A more detailed analysis reveals that the performance of VIL is at least as good as AL for all three task complexities ($\mathbf{RQ}_{1,2}$). For *easy* tasks VIL outperforms AL considerably. In early phases, the accuracy curve is very steep and converges at higher levels. Using iteration 20 as an example, M = 0.96 (CI = [0.95 – 0.98]) for VIL whereas the performance of AL only reaches M = 0.85 (CI = [0.82 – 0.88]). One explanation may be that VIL enables faster capturing instances of *all* classes. This can be seen as an indicator that human intuition (in combination with data visualization) may be useful to solve the cold start problem of AL approaches (AL even performs weaker than RB at start). The results of the performance comparison for medium and difficult task complexities are similar. VIL performs slightly better than AL, RB can compete with AL. In the difficult case after 20 iterations, VIL performs with M = 0.42 (CI = [0.38 – 0.45]) in contrast to AL (M = 0.35, CI = [0.33 – 0.38]). However, we ascertain that VIL shows higher variations (Figure 4 center, right). We draw the conclusion that, with VIL, human control over the labeling process is beneficial, but may also lead to weaker performances for individuals. In general, we ascertain that VIL can compete with the remaining strategies ($\mathbf{RQ}_1$), even for complex labeling settings ($\mathbf{RQ}_2$). The comparison of all strategies with the results of the ULoP indicates remaining potential. Even for the difficult condition, ULoP still achieved outstanding performance (M = 0.70 at iteration 20).

5.1.2 Comparison of VIL-Support Techniques

The second factor in $\mathbf{PART}_1$ was the comparison of the 4 VIL-support techniques. The core question is whether VIL-support techniques perform differently ($\mathbf{RQ}_3$). According to the 4×3 experiment conditions, we compare VIL-support techniques for the easy, medium, and difficult complexity level, Section 4.6 provides additional details.

Results In Figure 5, results of the four VIL-support techniques can be compared for the *easy* task. Thus, we now explicitly distinguish the performances of different VIL-support techniques. One finding that can be identified without effort is the substantially better performance of all four candidates compared to AL. Compared to results of more difficult tasks VIL strategies are closer to the ULoP. We assume that, given an *easy* task, a clear separation of classes in 2D eases the identification of instances, leading to high *accuracies* very quickly. We conclude that differences between the four VIL-support techniques are negligible.

Figure 6 shows the performances of the VIL-support techniques for the *medium* task complexity. Again, it becomes apparent that VIL-techniques are able to compete with AL strategies ($\mathbf{RQ}_1$), particularly at the start of the process. In early phases of the labeling process (e.g., iteration 10), we assess the best performance for the *Class Colors* technique (M = 0.73, CI = [0.71 – 0.75] in contrast to the *2D Colormap* (M = 0.68, CI = [0.64 – 0.73]), *Convex Hull* (M = 0.70, CI = [0.65 – 0.74]), and the *Butterfly Plot* (M = 0.70, CI = [0.66 – 0.73]). Another insight for late iterations is the good performance of the three techniques displaying classifier information (*Class Colors* M = 0.83, *Convex Hull* M = 0.85 and *Butterfly Plot* M = 0.84), the Convex Hull approach performs best and keeps track with AL (M = 0.85) ($\mathbf{RQ}_{1,3}$). In turn, the *2D Colormap* approach cannot compete with the remaining techniques (M = 0.82). Finally, the ULoP outperforms all remaining strategies substantially which shows that there is still room for improvements.

The situation is similar for the *difficult* labeling task (see Figure 7). All VIL-support techniques perform comparatively well at the beginning. Here, the *2D Colormap* (no model information provided) slightly outperforms the remaining techniques ($\mathbf{RQ}_3$). In the course of the process, the two shape-based techniques (*Convex Hull* M=0.61 and *Butterfly Plot* M=0.60) achieve the highest *accuracies*. Overall, the VIL strategies perform at least as good as AL and RB, AL again seems to perform weaker than RB at first. One explanation is that AL has problems to capture representatives of all classes (cold start problem) which is may be more severe for a complex tasks. To further investigate the cold start problem, we refer the reader to Figure 3 which compares the accuracy of AL and RB for substantially more iterations. Here, we identify a break-even point at approximately 50 instances where AL starts to outperform RB. The observation strengthens the rationale to combine the strengths of AL and VIL, using the latter for entirely unlabeled data. One final insight gathered from Figure 7 is the again outstanding performance of ULoP.

In summary, we identified that VIL can compete with AL ($\mathbf{RQ}_1$) especially in early phases of the process. In addition, we ascertain a slight tendency of VIL-support techniques with classifier visualization to outperform the *2D Colormap* technique ($\mathbf{RQ}_3$).

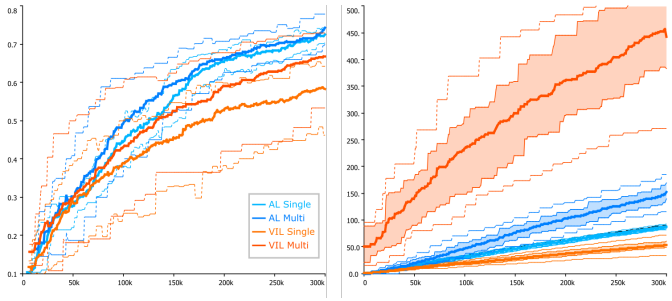


Fig. 8: Comparison of AL with VIL in combination with labeling single and multiple instances at once (**PART**$_2$). The duration of the labeling process is mapped to the x-axis ([ms] - 300k = 5 min). Left: AL achieves higher accuracies (y-axis) than VIL for both single and multi-instance labeling. Right: the number of labeled instances is mapped to the y-axis. Assigning a label for multiple instances at once works particularly fast for VIL. In 5 minutes users were able to label approximately 450 instances on average.

5.2 PART$_2$: Labeling Single vs. Multiple Instances

PART$_2$ considered the question of whether the process can be made more efficient when multiple instances are labeled at once. In particular, we were interested whether or not VIL can make the process more efficient (**RQ**$_4$). The four candidate interfaces for **PART**$_2$ are *AL single labeling*, *AL multi labeling*, *VIL single labeling*, and *VIL multi labeling*, Section 4.7 provides additional information about the experiment design.

Analysis of Effectiveness We analyze the progression of accuracy of the four candidates measured over time (see Figure 8, left). The duration (300,000 ms) of the labeling process is mapped to the x-axis. The analysis of the *accuracy* (mapped to the y-axis) provides three insights. First, in the beginning of the process the four candidates perform fairly balanced. Second, both AL strategies (blue colors) achieve higher *accuracies* in less *time* in the remaining phase of the process. Third, labeling multiple *instances* at once improves the efficiency. This accounts for AL as well as for VIL. Overall, with respect to the *accuracy* over *time*, VIL cannot compete with AL. One explanation of the lower *accuracy* values is associated with an observation we made several times in the study. Selecting and filtering a large number of identical instances requires a considerable portion of *time*. Thus, many participants did not label all of the classes (leading to weaker accuracies), as this was not subject of the task introduction.

Analysis of Efficiency We analyze the number of labeled instances over time for all four candidates (see Figure 8, right). We make two observations. First, the efficiency of *VIL multi labeling* is considerably better than the three remaining candidates. Thus, labeling multiple instances at once with a VIL-based interface can substantially increase the efficiency of the labeling process (**RQ**$_4$). We observed that users partially labeled 50 or more *instances* at once leading to a massive increase of efficiency compared to the other interfaces. Even the minimum performance of all users is considerably better than the maximum performance of any user using the remaining interfaces. Second, both AL-based candidates outperform *VIL single labeling*. This can be explained with the overhead of VIL approaches requiring additional exploration, identification, and selection of single candidate instances.

5.3 PART$_3$ User Strategies and User Feedback

In **PART**$_3$, we focus on the question whether or not users develop strategies for the selection of instances in VIL (**RQ**$_5$) and if so, how these strategies relate to VIL-support techniques and AL strategies (**RQ**$_6$). The experimental setup of **PART**$_3$ follows the description in Section 4.8.

5.3.1 User Strategies for Labeling Data

The observation of participants during **PART**$_{1\text{-}2}$ revealed a series of user strategies for selecting labeling candidates (**RQ**$_5$). We classified these strategies into data-centered and model-centered strategies in a joint discourse of the authors. Data-centered strategies focus on characteristics of data instances (elementary or synoptic level). Model-centered strategies are based on visual feedback of the current state of the classification model.

Data-Centered Strategies **Dense Areas First:** Collections of instances that form dense clusters are preferred during the labeling process. This supports the classification performance by learning the information for many instances at once. **Centroid First:** Special type of Dense Areas First. Instances that are at the center of clusters are labeled first, in order to assign labels to instances that are representative for a cluster. **Equal Spread:** The user tries to assign labels to instances that are well distributed (in 2D), to make sure that there are no areas in the original (high-dimensional) space that do not contain labeled instances. **Cluster Borders:** Instances that are at the border between two clusters are labeled, in order to give the classifier information that helps to better separate the clusters. **Outliers:** Outliers of the data set are labeled explicitly, in order to allow the classifier to learn about the range of instances that belong to one class. **Ideal Label:** The user only assigns labels to the instances that are ideal candidates or representatives of the respective class. The motivation for some users applying this strategy was based on data-semantical reasons.

Model-Centered Strategies **Class Distribution Minimization:** The spread of a class in the 2D representation (e.g., represented with a convex hull) is to be minimized. **Class Borders:** The user tries to achieve clearly separated borders between classes based on the visual feedback on spatial class distribution (e.g. based on the size of the convex hull). **Class Intersection:** The labeling process aims at minimizing possible ambiguities in the intersection between classes (e.g. depicted with overlaps of convex hulls). **Class Outliers:** Users label those instances that are assigned to a class but are far away from the class center of gravity. Referring to the convex hull, this can be identified with a spike.

5.3.2 Subjective User Feedback

Finally, we report on the subjective feedback we received in the concluding interviews and questionnaire. Table 1 provides an overview of average scores (5-point Likert scale, with 5 being 'very good'). The subjectively preferred techniques are Convex Hulls and Class Colors (**RQ**$_6$). Overall, information about the state of the classification model (Class Color, Convex Hulls, Butterfly Plot) was welcomed by most participants. We identified a shift in the labeling strategies towards model-centered characteristics, such as class distributions and class outliers, provoked by the additional visual encodings. One potential drawback of the shift towards model-centered information is neglecting data-centered properties as revealed by dimensionality reduction in combination with scatterplots.

In the interviews, we also received rich user feedback on the usability of the different VIL-support techniques (**RQ**$_6$). **2D Colormap:** Users remarked the disadvantage of missing model information. In turn, the simplicity of the interface was welcomed for the complex labeling task. **Class Color:** Users welcomed the direct feedback about the current state of the model by means of colors. For complex labeling tasks with many colors, users had problems in the distinction of some (categorical) colors. **Convex Hulls:** Many users liked the combination of color and shape-based information about the current state of the model. The distribution of classes in the 2D representation was easily comprehensible. However, overlays of many semi-transparent layers caused some problems in distinguishing classes. **Butterfly Plots:** The Butterfly plot obtained the most user feedback. Positive aspects are the

	2D Colormap	Class Colors	Convex Hulls	Butterfly Plot
Score	1.7	4.3	4.4	3,7

Table 1: User preferences on VIL-support techniques in **PART**$_1$ and **PART**$_2$. Convex Hulls achieved the highest scores, followed by Class Colors.

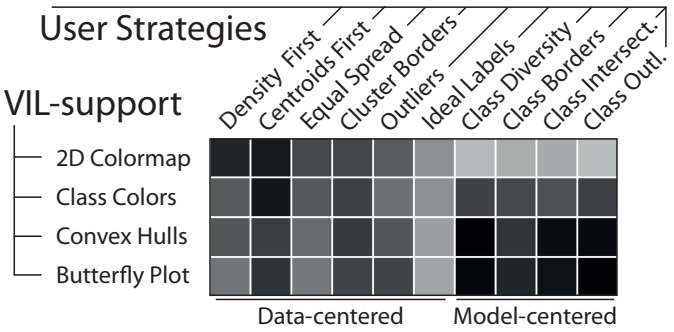


Fig. 9: Usefulness of VIL-support techniques for ten identified user strategies for labeling data. The strategies can be partitioned into data-centered and model-centered strategies. User preference is depicted with color values from gray (low) to black (high).

depicted information about the model including an indication of class outliers and the class centers, as well as the more compact fill area compared to convex hulls. However, many users were confused by the non-regular shapes that distracted them from the underlying data.

6 Discussion and Future Work

We conducted an experiment with three parts to assess the performance of VIL in comparison to AL. The results revealed several insights about the applicability of VIL-support techniques, but also shed light on current limitations. Based on the experiment parts, we experienced a series of user strategies for selecting data for labeling which provide interesting stimuli for future research. In the following, we highlight lessons learned from the experiment, discuss insights from the interviews, and point out future work.

6.1 Dimensionality Reduction

The interviews revealed that every user decided to conduct large parts of $\mathbf{PART}_1$ and $\mathbf{PART}_2$ with t-SNE as dimensionality reduction technique. Virtually all participants only used the remaining techniques for validation purposes, to have a second perspective. By design, t-SNE was the default technique and we cannot completely exclude that this may have caused a bias. Still, we can confirm that users unanimously argued that they sought for a technique that separates cluster structures best, according to the study conducted by Lewis et al. [34] showing that users think seeing a cluster is a sign of quality of the method. One drawback of the unanimous vote is that we cannot make statements about the performance of alternative techniques for dimensionality reduction. A systematic study on the benefit of different dimensionality reduction techniques on the labeling process is an open topic and subject of future work. Another point of discussion is class separability which became more difficult for complex labeling tasks in $\mathbf{PART}_1$. For strongly overlapping classes dimensionality reduction may not be the best choice, as the individual classes may not separate well in the resulting 2D mapping. One reason for this is that dimensionality reduction techniques do not consider class information. An alternative solution may be the use of Linear Discriminant Analysis (LDA) [20] or other supervised methods [69] for dimensionality reduction. These approaches take class information into account and thus may enable a more appropriate dimensionality reduction and visualization of the data.

6.2 Analytical Guidance

Our work fundamentally deals with the question on how to guide users, an important research challenge in visualization [26]. Considering data or model-centered perspectives, guidance can be achieved by different strategies coming either from AL (e.g. highlighting instances near decision boundaries) or VIL (e.g. visualizing cluster centroids). Designing optimal guidance models, however, is a challenging direction for future research. Summarizing $\mathbf{PART}_3$, we have identified three promising approaches for guidance that we plan to implement and evaluate in the future: (i) providing instant feedback on the (estimated) benefit of labeling a certain instance, plus the visualization of the development of accuracy over time; (ii) guiding the user in a way that the distribution of labels across classes becomes balanced to avoid biasing the learning algorithm towards a certain class; (iii) leveraging analytical class separability measures (e.g. Sips et al. [63]) to adaptively select a suitable visualization or to continuously select the best dimensionality reduction technique for the given dataset during the labeling process (i.e. the one which maximizes the class separability). A final issue of discussion is the study design with conditions building up on each other. While we observed a positive effect towards user guidance, we cannot preclude a certain bias from the VIL strategies in predefined order. An investigation of this potential bias is a subject of future work.

6.3 User-based Labeling Strategies vs. Active Learning

One goal of our experiment was the identification of 10 VIL strategies (cf. Section 5.3.1) ($\mathbf{RQ}_3$). A question that arises in this context is which similarities and differences between VIL and AL strategies exist and whether one approach can learn from the other. We observe for example that many VIL strategies have a direct counter part in AL, e.g. "Density First" corresponds to density-based sampling and "class intersection" is a variant of uncertainty sampling. Other strategies, however, are special to VIL, e.g., "Equal Spread" and "Outliers". Learning about VIL strategies developed by users may be a valuable source of information for novel AL strategies. Conversely, AL strategies may inspire novel visual guidance approaches for VIL. Furthermore, the list of VIL strategies may extend in future investigations and represents a topic for further investigation.

6.4 Upper Limit of Performance

The ULoP was much better than the AL and VIL strategies in all results. This shows that in both domains there is still potential for improvements which justifies future research. One exception was the easy labeling task examined in $\mathbf{PART}_1$ where some users achieved similar performance values for some early iterations (cf. Section 5.1). This leads to the question which guidance strategy best approximates the upper limit. In this context, it may be interesting to compare concrete candidate suggestions proposed by the upper limit of performance, those of AL strategies, and those of users.

6.5 Visual Instance Representation

In this study, we have employed handwritten digits data which is easy to visualize and self-explanatory as well. In general, proper visual representations of instances are needed to enable users grasp the data characteristics [7]. Promising classes of techniques addressing this challenge are visual identifiers like images of soccer players [3], glyph designs [8], visualizations showing the feature space [29], or visual-interactive solutions allowing to grasp detailed information about data on demand.

7 Conclusion

We examined the performance of visual-interactive labeling (VIL) strategies in comparison to active learning (AL) and random sampling. The overall objective was to assess whether or not VIL can compete with AL. We conducted an experiment with three parts, each with focus on a different aspect. First, we examined four VIL-support visualization techniques for three different task complexities and identified that convex hulls depicting the current model state are particularly suitable to support users in labeling data instances. In addition, we ascertained that all tested VIL-support techniques can compete with the performance of AL labeling strategies, at least in the examined first 50 labeling iterations where the cold start problem of AL is most severe. Furthermore, we identified that VIL outperforms AL for easy tasks and can keep up for more difficult labeling tasks. Second, we assessed the positive effect of VIL for assigning labels to multiple instances at once. While AL outperforms VIL with respect to effectiveness, VIL leads to a substantial increase in efficiency. Third, a reflection of the experiment including observation and interviews of participants revealed ten user-based data selection strategies that may form a promising basis for future VIL and AL approaches, e.g., to incorporate analytical guidance in the labeling process.

REFERENCES

[1] American Psychological Association. *Publication manual of the American psychological association (6th edition)*. American Psychological Association Washington, 2010.

[2] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *IEEE Visual Analytics Science and Technology (VAST)*, pp. 43–52, 2014.

[3] J. Bernard, C. Ritter, D. Sessler, M. Zeppelzauer, J. Kohlhammer, and D. Fellner. Visual-interactive similarity search for complex objects by example of soccer player analysis. In *Proc. of IVAPP, VISIGRAPP*, pp. 75–87, 2017. doi: 10.5220/0006116400750087

[4] J. Bernard, D. Sessler, A. Bannach, T. May, and J. Kohlhammer. A visual active learning system for the assessment of patient well-being in prostate cancer research. In *IEEE VIS Workshop on Visual Analytics in Healthcare (VAHC)*, pp. 1–8. ACM, 2015.

[5] J. Bernard, D. Sessler, T. Ruppert, J. Davey, A. Kuijper, and J. Kohlhammer. User-based visual-interactive similarity definition for mixed data objects-concept and first implementation. In *WSCG*, vol. 22. Eurographics, 2014.

[6] J. Bernard, M. Steiger, S. Mittelstädt, S. Thum, D. Keim, and J. Kohlhammer. A survey and task-based quality assessment of static 2d colormaps. In *Electronic Imaging, SPIE Conference on Visualization and Data Analysis*, vol. 9397, pp. 93970M–93970M–16. SPIE Press, 2015.

[7] J. Bernard, M. Zeppelzauer, M. Sedlmair, and W. Aigner. A Unified Process for Visual-Interactive Labeling. In M. Sedlmair and C. Tominski, eds., *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2017. doi: 10.2312/eurova.20171123

[8] R. Borgo, J. Kehrer, D. H. S. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics State of the Art Reports*, EG STARs, pp. 39–63. Eurographics Association, May 2013.

[9] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *IEEE Visual Analytics Science and Technology (VAST)*, pp. 83–92. IEEE, 2012.

[11] K. Charmaz. *Constructing grounded theory*. Sage, 2014.

[12] K. Chen and L. Liu. Clustermap: Labeling clusters in large datasets via visualization. In *ACM Conference on Information and Knowledge Management (CIKM)*, pp. 285–293. ACM, New York, NY, USA, 2004.

[13] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[14] A. Culotta and A. McCallum. Reducing labeling effort for structured prediction tasks. In *Conference on Artificial Intelligence (AAAI)*, pp. 746–751. AAAI Press, 2005.

[15] G. Cumming and S. Finch. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, 60(2):170, 2005.

[16] I. Dagan and S. P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Conference on Machine Learning (ICML)*, pp. 150–157. Morgan Kaufmann Pub., 1995.

[17] C. K. Dagli, S. Rajaram, and T. S. Huang. Leveraging active learning for relevance feedback using an information theoretic diversity measure. In *Conference on Image and Video Retrieval*, pp. 123–132. Springer, 2006.

[18] R. O. Duda, P. E. Hart, D. G. Stork, et al. *Pattern classification*, vol. 2. Wiley New York, 1973.

[19] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recogn. Lett.*, 30(1):27–38, 2009.

[20] R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x

[21] S. Garg, I. Ramakrishnan, and K. Mueller. A visual analytics approach to model learning. *na*, pp. 67–74, 2010.

[22] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(12):2839–2848, 2012.

[23] B. Höferlin, R. Netzel, M. Höferlin, D. Weiskopf, and G. Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *IEEE Visual Analytics Science and Technology (VAST)*, pp. 23–32, 2012.

[24] S. C. Hoi, R. Jin, and M. R. Lyu. Large-scale text categorization by batch mode active learning. In *World Wide Web*, pp. 633–642. ACM, 2006.

[25] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[26] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *IEEE Visual Analytics Science and Technology (VAST)*, pp. 3–10. IEEE, 2010.

[27] I. T. Jolliffe. *Principal Component Analysis*. Springer, 3rd ed., 2002.

[28] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 453–456. ACM, New York, NY, USA, 2008. doi: 10.1145/1357054.1357127

[29] J. Krause, A. Perer, and E. Bertini. Infuse: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Trans on Vis. & Comp. Graph.*, 20(12):1614–1623, 2014. doi: 10.1109/TVCG.2014.2346482

[30] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[31] N. Landwehr, M. Hall, and E. Frank. Logistic model trees. *Machine Learning*, 59(1-2):161–205, 2005.

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[33] J. M. Lewis, M. Ackerman, and V. R. de Sa. Human cluster evaluation and formal quality measures: A comparative study. In *Annual Meeting of the Cognitive Science Society (CogSci)*, pp. 1870–1875, 2012.

[34] J. M. Lewis, L. van der Maaten, and V. R. de Sa. A behavioral investigation of dimensionality reduction. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2012.

[35] G. M. Mamani, F. M. Fatore, L. G. Nonato, and F. V. Paulovich. User-driven feature space transformation. In *CGF*, pp. 291–299, 2013.

[36] N. A. H. Mamitsuka. Query learning strategies using boosting and bagging. In *Conference on Machine Learning (ICML)*, vol. 1. Morgan Kaufmann Pub., 1998.

[37] A. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In *Conference on Machine Learning (ICML)*, pp. 350–358. Morgan Kaufmann Pub., San Francisco, CA, USA, 1998.

[38] J. Möhrmann, S. Bernstein, T. Schlegel, G. Werner, and G. Heidemann. Improving the usability of interfaces for the interactive semi-automatic labeling of large image data sets. In *Human Computer Interaction. Design and Development Approaches*, pp. 618–627. Springer, 2011.

[39] F. Olsson. A literature survey of active machine learning in the context of natural language processing. *Technical Report, no. 3600*, 2009.

[40] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Mller, and H. Piringer. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 23(1):611–620, 2017.

[41] F. V. Paulovich, M. C. F. Oliveira, and R. Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 27–36. IEEE Computer Society, Washington, DC, USA, 2007.

[42] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1880–1897, 2009.

[43] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. C. North, and D. A. Keim. Human-Centered Machine Learning Through Interactive Visualization: Review and Open Challenges. In *Symp. on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[44] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 23(01):241–250, 2016.

[45] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Readings in Information Retrieval*, 24:5, 1997.

[46] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18(5):401–409, 1969. doi: 10.1109/T-C.1969.222678

[47] A. Sarkar, M. Spott, A. F. Blackwell, and M. Jamnik. Visual discovery and model-driven explanation of time series patterns. In *Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 78–86. IEEE, 2016.

[48] T. Scheffer, C. Decomain, and S. Wrobel. Active hidden markov models for information extraction. In *Conference on Advances in Intelligent Data Analysis (IDA)*, pp. 309–318. Springer-Verlag, London, UK, UK, 2001.

[49] T. Schreck and C. Panse. A new metaphor for projection-based visual analysis and data exploration. In *Visualization and Data Analysis 2007*, SPIE Proceedings, pp. 64950L–64950L–12, 2007.

[50] T. Schreck, M. Schler, F. Zeilfelder, and K. Worm. Butterfly plots for visual analysis of large point cloud data. In *WSCG*, pp. 33–40. University

of West Bohemia, Plzen, 2008.
[51] T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193, 2010.
[52] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):2634–2643, 2013.
[53] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, 2012.
[54] D. Seebacher, M. Stein, H. Janetzko, and D. A. Keim. Patent Retrieval: A Multi-Modal Visual Analytics Approach. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pp. 013–017. Eurographics, 2016.
[55] C. Seifert and M. Granitzer. User-based active learning. In *IEEE Conference on Data Mining Workshops (ICDMW)*, pp. 418–425, 2010.
[56] B. Settles. Active learning literature survey. Tech. Report 1648, Univ. of Wisconsin–Madison, 2009.
[57] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1467–1478. Computational Linguistics, 2011.
[58] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Empirical Methods in Natural Language Processing*, pp. 1070–1079. Computational Linguistics, 2008.
[59] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pp. 1289–1296, 2008.
[60] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proc. of the 5th Ann. Worksh. on Comput. Learning Theory*, COLT '92, pp. 287–294. ACM, New York, NY, USA, 1992.
[61] C. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.
[62] E. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.
[63] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.
[64] J. Stahnke, M. Drk, B. Mller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):629–638, 2016.
[65] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Selected Topics in Signal Proc.*, 5(3):606–617, 2011.
[66] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
[67] J. Vendrig, I. Patras, C. Snoek, M. Worring, J. den Hartog, S. Raaijmakers, J. van Rest, and D. A. van Leeuwen. Trec feature extraction by active learning. In *TREC*, 2002.
[68] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol.*, 2(2):10:1–10:21, 2011.
[69] Y. Wang, K. Feng, C. Xiaowei, J. Zhang, C.-W. Fu, M. Sedlmair, X. Yu, and B. Chen. A Perception-Driven Approach to Supervised Dimensionality Reduction for Visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2017. to appear.
[70] M. Ware, E. Frank, G. Holmes, M. Hall, and I. H. Witten. Interactive machine learning: letting users build classifiers. *Human-Computer Studies*, pp. 281–292, 2001.
[71] N. Weber, M. Waechter, S. C. Amend, S. Guthe, and M. Goesele. Rapid, Detail-Preserving Image Downscaling. In *ACM SIGGRAPH Asia*, 2016.
[72] Y. Wu, I. Kozintsev, J.-Y. Bouguet, and C. Dulong. Sampling strategies for active learning in personal photo retrieval. In *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 529–532. IEEE, 2006.