Active Learning for Image Recognition using a Visualization-Based User Interface

Christian Limberg*†, Kathrin Krieger*, Heiko Wersing†, Helge Ritter*
*Bielefeld University
Universitätsstraße 25, 33615 Bielefeld, Germany
CITEC, Neuroinformatics Group
Email: {climberg, kkrieger, helge}@techfak.uni-bielefeld.de
†HONDA Research Institute Europe GmbH
Carl-Legien-Straße 30, 63073 Offenbach, Germany
Email: heiko.wersing@honda-ri.de

***Abstract*—This paper introduces a novel approach for querying samples to be labeled in active learning for image recognition. By using dimension reduction techniques to create a 2D feature embedding for visualization, the user is able to efficiently label images for training a classifier. This is made possible by a querying strategy specifically designed for the visualization, seeking optimized bounding-box views for subsequent labeling. The approach is implemented in a web-based prototype. It is compared in-depth to other active learning querying strategies within a user study we did with 31 participants. With our approach the participants could train a more accurate classifier than with the other approaches on a challenging data set. Additionally, we demonstrate that due to the visualization, the number of labeled samples increases and also the label quality improves.**

***Index Terms*—Active Learning, Classification, Pattern Recognition, Image Recognition, Object Recognition, User Interface, Visualization, Dimension Reduction**

I. Motivation

In a classification task, there are machine learning models that can be trained incrementally and samples can be labeled stepwise by the user. Active learning [16] is an efficient training technique, where the samples, which are predicted to deliver the highest improvement for the classifier, are chosen for being labeled. There are several approaches for selecting the samples to be queried. However, it depends on the actual data which approach yields the best accuracy [18].

Having this in mind, we try to find a more efficient way for applying active learning. The common practice is to ask the human for a label of one single sample at a time [17]. Since this is a monotonous task and therefore often leads to mislabeled samples, we want to intervene already at this point by using a labeling user interface which is not only capable of boosting the performance of the classifier and increase the number of labeled samples, but also gives the human a more pleasurable experience while training the classifier. Another goal is to give the human a better idea about the inner representation of the trained model. This insight may lead to a better understanding where strengths and weaknesses of a feature representation are. To facilitate human labeling of high-dimensional samples, we use dimension reduction techniques to visualize the data in a 2D feature embedding space. We use this for improving active querying in an image recognition task.

There are some approaches towards machine learning using such a visualization. Recently, Cavallo et al. [1] introduced an approach for not only visualizing high dimensional data, but also changing both the data in the feature embedding space and in high dimensional space. For instance, after changing data in feature embedding space it can be explored what effect this has in the high dimensional data and vice versa. Iwata et al. [6] introduced an approach where the user can relocate the data in a visualization to be more representative for him. This can be useful if data is clustered in different categories and a category should be located in one region of the visualization space. It is also useful for ordering data, if it has a natural ordering like numbers or letters.

More related to active learning, there are approaches using scatter plots for visualizing data to facilitate labeling. Huang et al. [5] improved the labeling process of text documents showing the human visualizations of the feature space. The text data is visualized by t-SNE [15], force-directed graph layout and chord diagrams. Hongsen et al. [11] used semi-supervised metric learning to train a visualization of video data. In both approaches, the data is displayed next to the scatter plot for labeling. The querying of samples is done manually by the user, so there is no active learning strategy involved directly, which we want to accomplish for image recognition.

We introduce an active querying technique which utilizes the visualization and enables an efficient training by finding bounding-box views in the visualization for labeling. We show in our contribution, that especially for image data using a visualization is favorable and that using our adaptive interface together with the proposed querying method is more efficient than state-of-the-art approaches.

II. Active Learning

Active learning is an efficient technique for training a classifier incrementally. One variant of it is pool-based active learning, where the features $\mathbf{X}$ with labels $\mathbf{Y}$ are divided in an unlabeled pool $\mathbf{U}$ and a labeled pool $\mathbf{L}$. A querying function selects the most relevant samples from $\mathbf{U}$ to be labeled by an oracle, which is in most cases a human annotator. As the training progresses, samples from the unlabeled pool $\mathbf{U}$ are labeled and put in the labeled pool $\mathbf{L}$. Simultaneously, the classifier c is trained online with the new labeled samples.

There were many research contributions in the past proposing querying methods for high performance gain of the classifier. An often used approach is Uncertainty Sampling (US) [8], originally proposed by Lewis et al. [10]. In US the classifier's confidence estimation c_p of the samples from the unlabeled pool are used to select those with the lowest certainty for querying: $argmin_{u \in \mathbf{U}} c_p(u)$. Another technique is query by committee (QBC) [9], [19], where the query is chosen that maximizes the disagreement of the classifiers. In our evaluation we use the vote entropy for measuring the disagreement of classifiers: $argmax_{u \in \mathbf{U}} - \sum_i \frac{V(y_i)}{C} log \frac{V(y_i)}{C}$ where y_i is a particular label and $V(y_i)$ is the number of classifiers voted for this label, C is the number of classifiers in the committee. In our evaluation we chose a linear Support Vector Machine, a Decision Tree and Logistic Regression as a committee of diverse classifiers.

III. Dimension reduction for visualization

There are many dimension reduction approaches to visualize a high-dimensional feature space in lower dimensions. Their training is usually unsupervised. An early approach is Principal Component Analysis (PCA) [4], where a small set of linearly uncorrelated variables having the highest variance in the data, called principal components, are extracted. Multidimensional Scaling (MDS) [21] is a technique for dimension reduction, which preserves the spatial relation of the high-dimensional data in the lower-dimensional space. A Self Organizing Map (SOM) [7], introduced by Kohonen in 1982, can be used for dimension reduction. By applying competitive learning SOMs can preserve topological properties in the lower dimensional map. In 2008, van der Maaten et al. proposed t-SNE [15], which is a variant of Stochastic Neighbor Embedding (SNE) [3]. By modeling data points as pairwise probabilities in both the original space and the embedding, using a gradient decent method to minimize the sum of Kullback-Leibler divergences, it is possible to create an embedding of high quality. Especially if there are classes with different variances in high dimensional space, t-SNE delivers reasonable results. Our preliminary experiments also show, that t-SNE is delivering best results compared to PCA and MDS for image data where classes consist of objects showed from different viewing positions, like in the OUTDOOR data set [14] that we will also use in our evaluation. Because of these advantages, we use t-SNE as a dimension reduction technique in our experiments, but basically every other dimension reduction approach can be used as well.

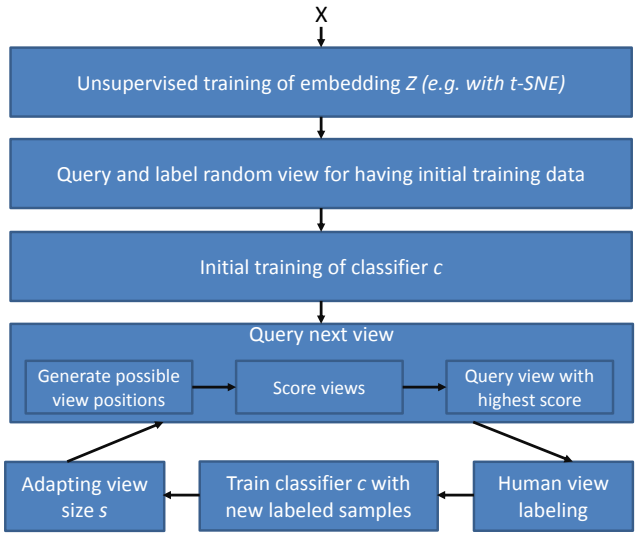

Fig. 1. General workflow diagram describing active learning with visualization.

IV. Adaptive Visualization View Querying (A2VQ)

The underlying idea is to query the samples within a bounding-box view of the visualization which we denote as a view $\mathbf{v}$. The goal of our approach is to query the optimal view for labeling of its enclosed samples.

In the following we introduce the Adaptive Visualization View Querying (A2VQ) approach for querying in active learning using an adaptive visualization. The overall workflow is illustrated in Fig. 1. At first, we use the t-SNE algorithm to reduce the dimensions of data set $\mathbf{X}$ to 2D for visualization. We normalize the output by applying feature scaling so that values of each of the two dimensions are between 0 and 1, naming this normalized embedding feature space $\mathbf{Z}$. In the following we refer $\mathbf{Z}_i$ as the visualization of sample $\mathbf{U}_i$.

Since we assume to have no label information at the beginning, active training starts with an empty $\mathbf{L}$. So labeling of one or more random views by the human is necessary to initially train a classifier for our approach. Then confidences for samples of $\mathbf{U}$ are calculated by the classifier, used to query the optimal view (described in detail in the next chapter). The queried view can be labeled e.g. by a user with our proposed user-interface. Then the classifier is trained incrementally with the newly labeled samples. After this training epoch, a new optimal view is queried with the retrained classifier and the process repeats.

We think a querying method is necessary for an efficient labeling because a visualization of more complex data sets

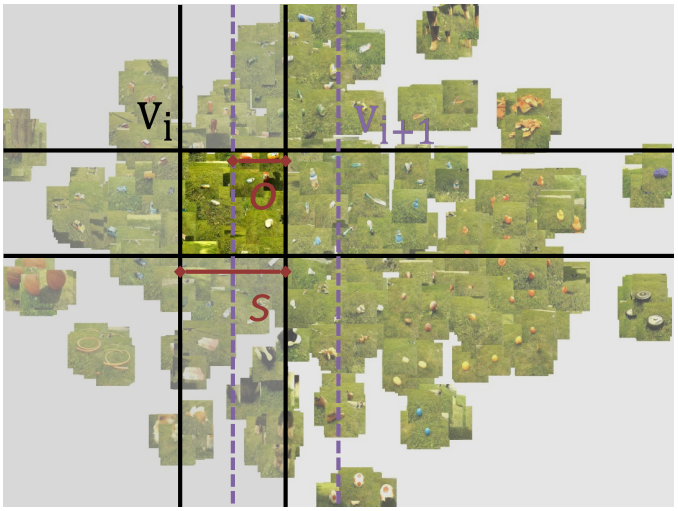

Fig. 2. t-SNE visualization of 50 objects from the OUTDOOR data set with illustrated sliding window approach. In one iteration of sliding window, all views of the visualization are scored by A2VQ's scoring function. The possible views are generated by moving the squared template with side length s in overlap o steps from the upper left to the bottom right corner. The view with highest score is queried for labeling and displayed in our web-based user interface.

Fig. 3. Querying user interface showing a view queried by A2VQ. The user can label samples by selecting their thumbnails by dragging rectangles in the visualization. The class name can be entered in an input formula. There are certain possible strategies for labeling, like label everything, label only the biggest clusters or label only outliers. With a click on the button *Query next view* the classifier is retrained with the new labeled samples and a new view is queried with A2VQ.

can be confusing for the human as there are too many classes and the images are highly overlapping as one can see in Fig. 2. Also we want to be able to actively query the samples which the classifier demands for efficient training.

A. Visualization View Querying

To query the optimal view we use a sliding window technique to cycle through a grid of possible bounding-box views that arise from a view size s and overlap amount o. The first view is positioned at $(0, 0)$ in $\mathbf{Z}$. By shifting the square $s-o$ in each dimension (illustrated in Fig. 2), there is a total number of $(1+\frac{1-s}{s-o})^2$ views to be evaluated. We therefore calculate a scoring function $r(\mathbf{v})$ for each view:

$$r(\mathbf{v}) = \frac{\sum_{u \in \mathbf{U_v}} (1 - c_p(u))}{m} \tag{1}$$

where $\mathbf{U_v}$ are the samples lying in the particular view, $c_p(u)$ is the classifier's confidences of the most certain class for sample u and m is the number of samples in the view with the most enclosed samples. By dividing by m not only the classifier's confidences of the view's samples are taken into account, but also the number of samples in the view. We do this for not querying views with few outlier samples with low confidences, as they can occur for instance at border areas in a t-SNE visualization (see Fig. 2).

After calculating r for each view generated by the sliding window approach, the view with the highest score r is queried for labeling.

B. User interface

The samples of the optimal view can be labeled with our user interface, also available at github[1] together with all implemented querying techniques. By applying an affine transformation the view is shown in full size with the corresponding sample images as scatter plot symbols. The resulting display is shown in Fig. 3. Due to the visualization most neighboring samples will receive the same label. Interactive selection techniques (Fig. 3) allow economic labeling of the samples within the view.

C. Adaptive view size

In addition to querying the best view for labeling, there is the question of finding the best view size s. A small s would not be efficient for labeling and a too large s makes it impossible for the human to recognize the images because there are too many. We investigated two heuristics for finding a suitable view size.

Number of Classes: In this heuristic we assume that showing the user about $c = 3$ different classes within a view results in best usability. We incrementally increase or shrink s we use a heuristic that is evaluated after each labeled view:

$$s = s + \sigma(\lambda * (c - n)) - 0.5 \tag{2}$$

where λ is the learning rate, n are the number of individual classes in the view after removing outlier classes

[1] https://github.com/limchr/A2VQ

with less than 5 samples and σ is the sigmoid function. Using the learning rate inside the sigmoid function, which is centered vertically by subtracting 0.5, enables us to incrementally change the view size to match c.

Preliminary (automated) experiments showed, that adjusting view size with upper heuristic converges to a proper view size with $\lambda = 0.05$. However, in our automated experiments we assumed that the user has perfect ability in labeling the samples and that he labels all samples within a view. But we train also ambiguous objects in our user study and so we want to give the human the change to skip samples. Since we can not evaluate n, we used another heuristic for choosing a view size:

Number of Samples: We assume that a view should not have more than $b = 100$ samples so that the user is able to recognize them while using our label interface. To determine the s that fits this assumption, we count the number of samples within all possible views. We sort this array in descending order and choose the highest 20% for calculating a mean, naming it m. We do this for several view sizes $\{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$ and choose the view size with the minimum $|b - m|$. In our user study we evaluated $s = 0.25$ and chose $o = 0.5s$. A smaller overlap would be possible but requires longer calculation time because more views have to be evaluated while querying.

V. Evaluation

A. Experiment

We did a user study for comparing A2VQ to the baselines US, QBC and random querying (RAND).

1) Participants: 31 participants (16 males, 13 females, 2 others) joined the study. The median of their age was 28 years. Most of the participants were students (27 students, 2 employed, 2 others). The participants were paid 5€ for completing the whole study which took 30 to 45 minutes. Three of the participants refused the money. The protocol was approved by the Bielefeld University Ethics Committee.

2) User interfaces: In the study participants labeled images with two different user interfaces. The first one was the already described user interface for A2VQ (see Fig. 3). Participants had to drag rectangles to mark multiple images with the same object in one view. Afterwards they had to choose the corresponding label from the drop down menu in the top left. To label the selected images with the chosen label, they had to click the *Label Selection* button. If the selection was wrong, they could click the *Remove Selection* button. Participants were told, that it is not necessary to label all images within one view because we wanted to give them the ability to skip samples in all approaches. If none of the images within a view could be labeled, the view with the next higher score was displayed. Otherwise, to go to the next view, participants had to click the *Next View* button in the top right.

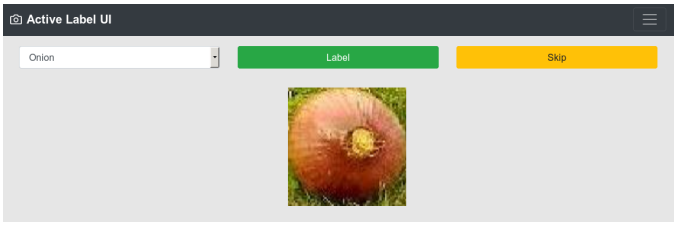

Fig. 4. Classic labeling interface for approaches US, QBC and RAND. The user also can skip images if he can not give a label.

The second user interface was used for labeling with US, QBC and RAND (see Fig. 4). To label an image, participants had to choose a label from the upper left drop down menu and click the *Label* button. If they were not sure about the label of an image, they could click the *Skip* button. After skipping an image we use DBQE [13] to prevent the querying of similar ambiguous images again, to speed up training.

3) Data set: We chose to use the OUTDOOR data set [14] for labeling in the experiment. The data set consists of 5000 images showing objects of 50 classes in a garden environment. Since this were too many classes to be labeled properly within a feasible time, we decided to reduce the data set to only seven classes. To make the labeling challenging for the participants, we selected object classes which might look very similar: *Onion*, *Orange*, *Potato*, *RedApple*, *RedBall*, *Tomato* and *YellowApple*. As a pre processing step, the objects are cropped using a color segmentation. For feature extraction we used the penultimate layer of the VGG19 deep convolutional net [20] trained for the imagenet competition, resulting in a 4096 dimensional feature vector. For evaluation we used a 80/20 train-test split. The test images are used to evaluate the classifier's performance. The images of the train set were presented in the user interfaces and labeled by the participants. We have chosen a 1 nearest Neighbor classifier with the same parameters for all the approaches. For estimating classifier confidences c_p we chose relative similarity [12]. The classifier is trained in an online fashion after each labeled image in the classic labeling interface or after each labeled image batch in A2VQ.

4) Questionnaires: Since we did not want to focus on a questionnaire analysis in our study, we decided to integrate just a short questionnaire in our study. The goal was to check whether our new approach delivers an idea about the inner representation of a trained model and might give a better understanding of strength and weaknesses of feature representations. We added, therefore, two questions:

Q1) Completing the task gave me an idea of the inner representation of a trained model.

Q2) Completing the task gave me a better understanding of strength and weaknesses of a machine learning model's feature representation.

Participants had to answer on a Likert scale from one (strongly disagree) to seven (strongly agree). They had

also the option to answer with N/A.

5) Task and procedure: At the beginning participants signed an informed consent. They read the global task instructions telling them that the main task is to label images to train a service robot to distinguish objects. Afterwards, they performed four experimental trials. Before a trial they had to first read specific task instructions. The instructions contained information about which of the two label interfaces they will use in the following trial and how to interact with it. The instructions did not provide any information about the underlying active learning approach. If a participant had to perform the label task with one of the two label interfaces for the first time, the experimenter showed a short video about the interaction with the label interface. In a trial, participants labeled images using the corresponding label interface. They had to be as fast as possible but also as accurate as possible. After five minutes the trial was stopped automatically by the system. After the trial, participants filled out a questionnaire.

6) Data recording: Whenever a participant labeled an image with any of the tested approaches, there were several information saved by the system:

- the participant's id
- the time in milliseconds since the start of the experiment
- the index of the labeled image
- the given label
- the ground-truth label
- the classifier's 0/1 accuracies on both train and test set

7) Experimental design: The experiment had a within subjects design, meaning that each participant labeled with each approach once. This resulted in four experimental trials, in which participants had to label the same images. It was, therefore, possible that participants became familiar with the images and improved their labeling performance during the experiment. To avoid such effects having an impact on the analysis, we varied the order of the experimental trials between participants. There exists 24 different possibilities to order the four experimental trials. Since we had 31 participants we needed 7 additional orders which were chosen randomly. The resulting 31 orders were randomly matched to the participants.

B. Analysis

Statistical tests were conducted with IBM, SPSS Statistics, Version 23.

1) Questionnaire: We extracted the answers for questions Q1 and Q2 from the participants' questionnaires. Higher values for Q1 indicate that completing the task delivered a better idea of the inner representation of the trained model. Higher values for Q2 provides a better understanding about strength and weaknesses of the model's feature representation. Since all the answers were given on a seven point scale, the values range between one and seven. We performed a two-sided Friedman's test (with $\alpha = 0$) for both questions each to check whether there are differences depended on the querying approach.

2) Recorded data: We investigated the impact of the four querying approaches A2VQ, US, QBC and RAND on three different parameters. The first parameter is the *classifier's acccuracy* for the test data set. Here the temporal progress of the accuracy and also the final accuracy after 5 minutes of training was explored. The second analyzed parameter was the *human label quality* which describes, how much of the data was labeled correctly by the participants. Finally, we analyzed whether the querying approaches have an impact on the *number of samples* which are labeled during the 5 minutes.

We aimed at analyzing whether there are significant differences between the different approaches in the three above mentioned parameters. Therefore, we first checked whether the data meets the assumptions to perform an ANOVA with repeated measures. By inspecting a boxplot, we noticed that all three parameters' data showed outliers. Furthermore the data were not normally distributed as assessed by Shapiro-Wilk's test ($p < .05$). According to this, we performed a two-sided Friedman's test (with $\alpha = .05$) instead of the ANOVA. For each of the three parameters, which showed significant results in Friedman's test, we checked whether there are significant differences among the approaches. Hence, we conducted multiple comparisons with a Bonferroni correction.

C. Results and discussion

Table II presents all means, medians and standard deviations of the analyzed data. Results of Friedman's test are summarized in Table III.

1) Questionnaire: In Fig. 5 top, the mean values of the questionnaire answers are plotted showing the lowest values for A2VQ. This means, that participants were most satisfied with the system usability in A2VQ. Additionally, this approach delivered them a better idea of the inner representation of a trained model than the baseline approaches. Finally, A2VQ gave them a better understanding of strength and weaknesses of a machine learning model's feature representation than other approaches. Anyway, Fig. 5 bottom shows that the differences between the approaches are very small compared with the distribution of the data.

A Friedman's test did not show any significant results for the three parameters.

2) Classifier's accuracy: Figure 6 shows the temporal progress of the classifier's accuracy on the test data during training. A2VQ has a slower increase of accuracy in early training while having a higher accuracy at the end (4.8% better than US). The slow rise might be because labeling with A2VQ is comparable with a depth-first search in a tree, while the other approaches are rather comparable with a breadth-first search, having a representation of each object class early in training. Most of the time QBC is performing better then US, which is performing better

TABLE I
NUMBER OF VALID ANSWERS, MEANS AND STANDARD DEVIATIONS FOR THE ANSWERS IN THE QUESTIONNAIRE.

	A2VQ			US			QBC			RAND		
	N	M	SD	N	M	SD	N	M	SD	N	M	SD
Q1	28	4.57	1.59	28	4.39	1.70	29	4.52	1.71	29	4.41	1.65
Q2	28	4.68	1.79	29	4.24	1.85	29	4.52	1.98	29	4.59	1.84

Note: The median for all questions and all approaches was $Mdn = 5$. A higher value means a higher agreement.

TABLE II
MEANS, MEDIANS AND STANDARD DEVIATIONS OF THE ANALYZED PARAMETERS.

	A2VQ			US			QBC			RAND		
	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
classifier's accuracy in %	77.52	81,54	15.83	72.73	75.15	13.21	71.61	73.08	13.81	68.41	70.00	14.49
human label quality in %	83.06	84.68	11.97	76.31	76.74	10.87	78.65	80.70	9.15	79.05	81.25	11.32
number of labeled samples	434.65	469	87.60	49.71	49	16.11	52.16	57	13.01	58	61	14.01

TABLE III
RESULTS OF FRIEDMAN'S TEST

	$\chi^2(3)$	p
Q1	1.965	.580
Q2	3.296	.348
classifier's accuracy in %	10.869	.012*
human label quality in %	9.311	.025*
number of labeled samples	60.650	<.001*

Note: An asterisk marks significant differences between the querying approaches on a level of $\alpha = .05$. Q1 and Q2 were the responses to questions of our questionnaire.

than RAND. All baseline approaches start to converge near the end of the experiment.

Friedman's test comparing the accuracies of the different approaches after five minutes training showed significant results. Post hoc tests reveal significant differences between A2VQ and QBC with p=.021 and between A2VQ and RAND with p=.002. This implies A2VQ delivers a better accuracy than RAND and QBC after five minutes training. Even if we did not find any significant differences between A2VQ and US, we can state that in our study A2VQ had the best mean accuracy compared with the other approaches after training the classifier for five minutes (see Table II).

3) Human label quality: In Fig. 7 a confusion matrix is displayed showing the true labels and the labels given by the participants averaged over all approaches. The labeling task was challenging for the participants who were not perfect oracles while labeling. This is especially noticeable at classes *RedApple*, *RedBall* and *Tomato* with a label quality of 80% and below.

To compare the label accuracy of the participants between the tested approaches, we performed Friedman's test. The test revealed significant results and, therefore, we performed multiple comparisons with a Bonferroni correction. This resulted in significant differences between A2VQ and the baseline approaches (A2VQ and US with $p = .005$, A2VQ and QBC with $p < .021$, A2VQ and RAND with $p = .030$). Figure 8 demonstrates the results. A2VQ has the best label quality, which is around 4% better than the second best (see Table II). The reason for this may be an improved human capability to see the objects in context with similar other objects and then to decide. Furthermore the RAND querying approach results in the second best label quality. This may lead to the assumption that classifier's uncertainty, which is used in US and QBC to query the most uncertain samples, is related to human uncertainty. Another interesting insight is, that even with a worse mean labeling quality, using US and QBC resulted in a better performing classifier than RAND (see Fig. 6 and Table II).

4) Labeled samples: Figure 9 shows how many samples were labeled within five minutes in the different experimental trials. The figure indicates, that people could label more samples using A2VQ while the number of labeled samples of the baseline approaches were comparable. The result of the statistical tests confirmed this observations. This outcome is as expected, because in A2VQ people can

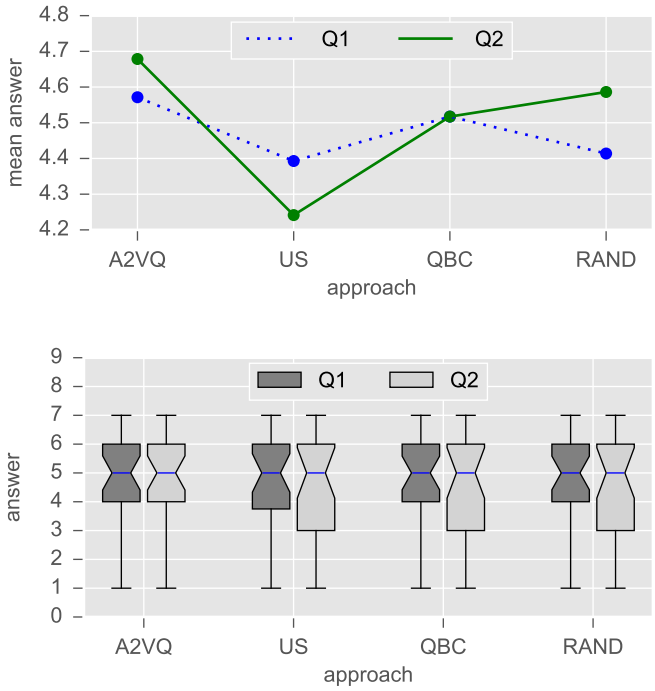

Fig. 5. Questionnaire answers of all 4 approaches. Higher values indicate stronger agreement. *Top:* Mean answers over all participants. *Bottom:* Box plots show the distribution of answers and the median as blue horizontal line.

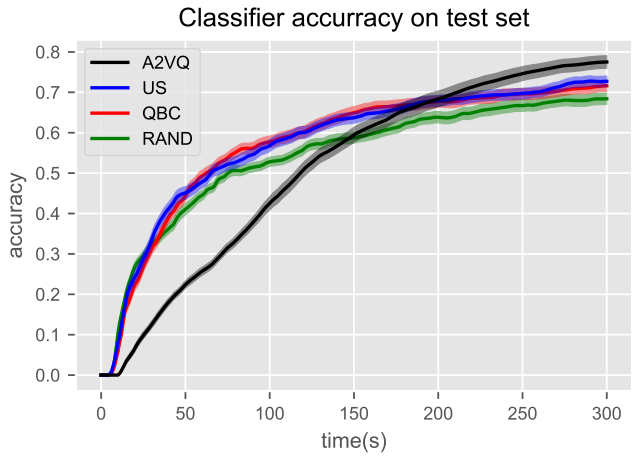

Fig. 6. Classifier's accuracy on held out test set while active training.

label multiple images with the same label while in baseline approaches just one image can be labeled at a time. Additionally, there was a significant difference between US and RAND.

VI. Conclusion

In this paper we have proposed to use dimension reduction techniques for applying active learning with a visualization. Therefore we introduced the querying approach A2VQ which queries optimal views for labeling by the

Fig. 7. Confusion matrix of human labels from all compared querying approaches.

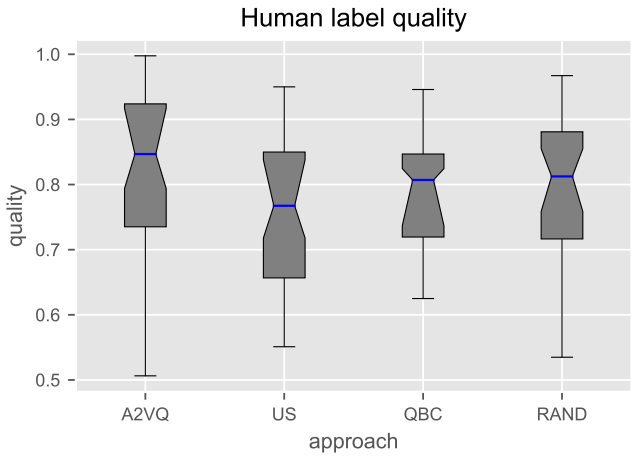

Fig. 8. Human label quality for tested approaches.

user. For showing it to the user, we developed a user interface, which we also evaluated in a user study. The study showed that using A2VQ the classifier's accuracy, the number of labeled samples and also the label quality improves compared to US, QBC and random querying.

There are many possible interesting further research in this topic. The user study showed that baseline methods have the advantage to faster respond at start of training. When training samples that can be ambiguous, we showed that the used DQBE [13] approach has a huge impact in boosting the speed by querying only meaningful samples. However, our study shows that after 100 seconds the fast increase in accuracy of the baseline methods saturates. So it may be worth to evaluate a hybrid model, that

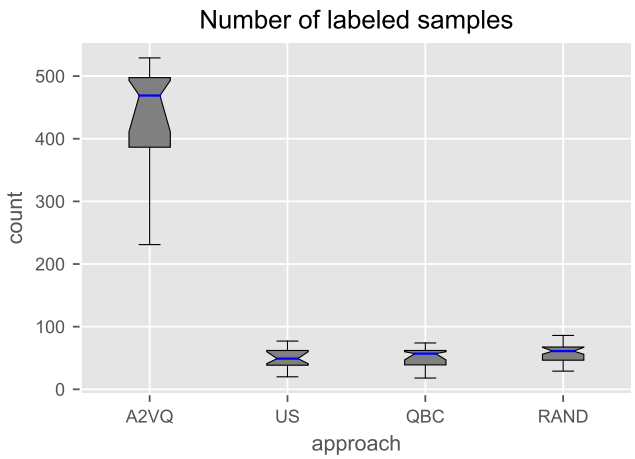

Fig. 9. Number of labeled samples of the different approaches.

first uses a classical active querying technique to query a few samples of each class for a fast learning of an initial classifier and then use A2VQ to label in depth. Using A2VQ also results in a higher label quality, as our study shows, so it may also correct former contradictions in labels, since we think that seeing patterns in contrast to other patterns facilitate to give the correct label.

It may is possible to use semi-supervised dimension reduction techniques [22] for a better visualization. Then, after each trained view not only the classifier is retrained but also the visualization is regenerated with new label information.

In the near future we will integrate A2VQ together with the labeling interface within a service robot [2], which interacts in a smart lobby environment. By showing the user interface on the robot's front touch screen we want to allow the user not only to teach the robot objects by a finger swipe, but also give him a feeling what the robot's internal representation of the objects might be.

References

[1] Marco Cavallo and Çagatay Demiralp. A visual interaction framework for dimensionality reduction based data exploration. In *Conference on Human Factors in Computing Systems CHI*, page 635, 2018.

[2] Stephan Hasler, Jennifer Kreger, and Ute Bauer-Wersing. Interactive incremental online learning of objects onboard of a cooperative autonomous mobile robot. In *International Conference on Neural Information Processing*, pages 279–290, 2018.

[3] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems (NIPS)*, pages 857–864, 2003.

[4] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[5] Lulu Huang, Stan Matwin, Eder J. de Carvalho, and Rosane Minghim. Active learning with visualization for text data. In *ACM Workshop on Exploratory Search and Interactive Data Analytics*, pages 69–74, 2017.

[7] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[8] Daniel Kottke, Adrian Calma, Denis Huseljic, Christoph Sandrock, George Kachergis, and Bernhard Sick. The other human in the loop - A pilot study to find selection strategies for active learning. In *International Joint Conference on Neural Networks IJCNN*, 2018.

[9] Bartosz Krawczyk and Michal Wozniak. Online query by committee for active learning from drifting data streams. In *International Joint Conference on Neural Networks IJCNN*, pages 2120–2127, 2017.

[10] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *International Conference on Research and Development in Information Retrieval (ACM-SIGIR)*, pages 3–12, 1994.

[11] Hongsen Liao, Li Chen, Yibo Song, and Hao Ming. Visualization-based active learning for video annotation. *IEEE Trans. Multimedia*, 18(11):2196–2205, 2016.

[12] Christian Limberg, Heiko Wersing, and Helge Ritter. Efficient accuracy estimation for instance-based incremental active learning. In *European Symposium on Artificial Neural Networks (ESANN)*, pages 171–176, 2018.

[13] Christian Limberg, Heiko Wersing, and Helge Ritter. Improving active learning by avoiding ambiguous samples. In *International Conference on Artificial Neural Networks (ICANN)*. Springer, October 2018.

[14] Viktor Losing, Barbara Hammer, and Heiko Wersing. Interactive online learning for obstacle classification on a mobile robot. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2015.

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.

[16] Maria Eugenia Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: an empirical study of common baselines. *Data Mining and Knowledge Discovery*, 31(2):287–313, 2017.

[17] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[18] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1070–1079, 2008.

[19] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Conference on Computational Learning Theory (COLT)*, pages 287–294, 1992.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[21] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[22] Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. Semi-supervised dimensionality reduction. In *SIAM International Conference on Data Mining*, pages 629–634, 2007.