

Active Learning for Multi-Label Image Annotation*

Mohan Singh, Eoin Curran, Pádraig Cunningham
University College Dublin

Technical Report UCD-CSI-2009-01
January, 2009

Abstract

Active learning is useful in situations where labeled data is scarce, unlabeled data is available and labeling has some cost associated with it. In such situations active learning helps by identifying a minimal set of items to label that will allow the training of an effective classifier. Thus active learning is appropriate for annotation tasks in multimedia, particularly in image labeling. In this paper we address the challenge of using active learning for *multi-labeling* of images in personal image collections. Multi-label learning covers situations where objects can have more than one class label and a learner is trained to assign multiple labels simultaneously. In this paper we report results on a learning system for labeling personal image collections that is both active and multi-label. The focus of the research has been to reduce the overall number of images that are presented to the user for labeling.

1 Introduction

Without mechanisms to annotate images we run the risk that our personal image collections become ‘write-only’ memories of our past lives. Recent studies [1, 2] have shown that users tend to ignore their past image collections because searching through the image collections involves a lot of time as the user still relies on examining individual image thumbnails in the image browser. These studies have also shown that the users are willing to spend extra effort annotating the images while storing them, making later access far easier. In this paper we present a multi-label active learning system that is specially designed to support the annotation of personal image collections. The main goal in the design of the system has been to minimize the overall number of images that the user is required to label in the training of the system.

The scenario that motivates this work is as follows. The user of the system loads a collection of images (say 50 - 100) into the system. The user wishes to attach labels from a collection of labels to these images – in the evaluation presented here there are four labels under consideration. The user *primes* the system with a small number of examples and counter-examples for these labels. In the normal active learning scenario these labels would be handled one at a time, i.e. the user would be queried for the label for the most informative image for each label in turn. In our system the active learning for all labels is done together, with images being selected for manual annotation that should be useful for the majority of learners. Clearly, these images will not be maximally informative for any one label classifier. However, they should be useful for most classifiers and should reduce the overall number of images presented to the user. The component classifiers are support vector machines (SVMs) and images are selected for manual annotation based on uncertainty where closeness to the

*The work was part supported by Science Foundation Ireland Grant No. 05/IN.1/I24.



Figure 1: Sample images from the Barcelona dataset

SVM margin is the measure of uncertainty. For the multi-label scenario the distances from the margin are converted to a probability and the probabilities across all SVMs are averaged to identify images that will be most useful.

In section 2, we present an overview of the task of labeling image collections and the dataset used in this paper. And in section 3 we review other research on active learning for multi-label image annotation. In the evaluation section 4 we first perform a cross-validation analysis to assess what classification accuracy is achievable on these image labels without any active learning. This acts as a baseline for the rest of the study. We then assess the performance of simple active learning for the four image labels. Then we evaluate the performance of multi-label active learning on all the labels taken together. The evaluation shows that our multi-label active learning strategy does reduce the overall number of images that the user has to label.

2 Annotating Personal Digital Image Collections

Image annotation, also known as image tagging, is a process by which labels or tags are associated with images, either manually, automatically or semi-automatically. A tag is a relevant keyword assigned to an image that will aid in the retrieval. The tags may come from a controlled ontology or users may be able to create their own tags. Image annotation for a personal image collection on a small scale can be done manually, but as the scale and size of image collections increase, the image annotation problem quickly becomes non-trivial. The problem becomes even more complex for large scale image databases for specialized fields such as medical imaging or space exploration.

Manual annotation though time consuming and laborious provides a description of an image at the right level of abstraction. But it can be a highly subjective task, because we as individuals interpret images in different ways. Automatic or semi-automatic image annotation based on low level visual features can be relatively fast but visual features alone cannot capture all concepts in an image that might be useful for annotation. The *semantic gap* refers to the difference between these low level features (see section 4) and the high level description needed for image interpretation. It is an active research topic in image annotation and many publications recently have focused on reducing the semantic gap in order to improve annotation performance of image tools [3, 4, 5].

2.1 Multi-label Image Dataset

In many annotation tasks, such as the annotation of personal image collections, multiple labels may apply to a single image. Figure 1 below shows four sample images from the dataset ¹ used in this paper. The dataset is composed of urban scenes from Barcelona. The example images were annotated by human with the following tags: ‘Buildings’, ‘Flora’, ‘People’ and ‘Sky’.

The dataset consists of 139 urban scene images and four overlapping labels: ‘Buildings’, ‘Flora’, ‘People’ and ‘Sky’. Each image has a minimum of two tags and each label is present in at least 60

¹The image dataset can be downloaded at (<http://mlg.ucd.ie/content/view/61>)

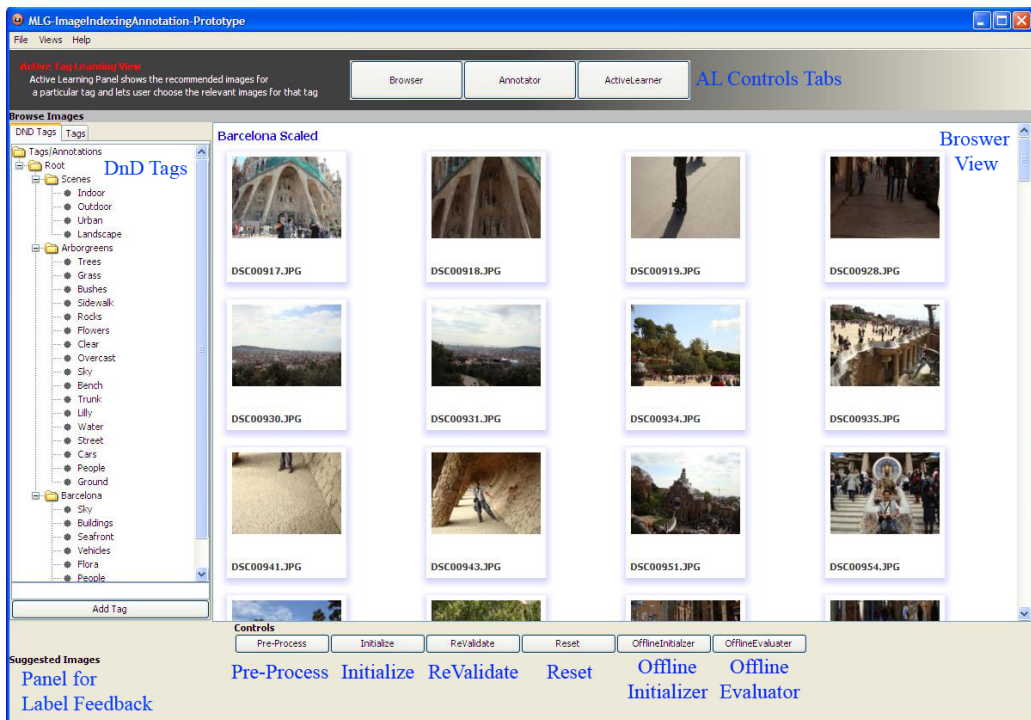


Figure 2: Image Annotation Prototype v1.0

images.

2.2 Prototype System for Image Annotation

A prototype system has been developed as part of this work (see figure 2) that allows the user to manually annotate a seed set of images and then interact with an active learning system in labeling the remaining images. The process starts off with the preprocessing of image data in order to extract low-level features, and the selection of a subset of these features to use in classification. Then the user initiates the semi-automatic labeling process by identifying a small set of positive and negative examples for each label. The system trains SVMs for each of these labels using these seed samples, and the multi-label active learning process selects unlabeled images from the pool of unlabeled examples to present to the user for labeling. The user assigns the appropriate labels to these query images and the SVMs are retrained with this new information. If the user does not attach a label to a query image, it is assumed that the image is a negative example for that label.

3 Active Learning for Image Annotation

This image labeling problem has all the characteristics of a situation where active learning is appropriate: labeled data is scarce, there is an available pool of unlabeled data and labeling expertise is available but labeling effort should be minimized. This idea of using active learning to assist in image labeling has received a lot of research attention [6, 7, 8] and we wish to build on that work to apply active learning in a multi-label scenario for the image annotation problem.

The standard active learning scenario is shown in Algorithm 1. The learner receives a small set of labeled examples D and a pool of unlabeled examples X . The *active* step is 2(b) where the user

is asked to label one of the unlabeled examples. Thus the key step in determining the effectiveness of the AL process is step 2(a) where the example to present to the user is selected.

Algorithm 1: Active Learning with Selective Sampling

$\mathbf{AL}(X, D, f())$: D is labeled set, X is unlabeled set, $f()$ is labeling function

1. $h \leftarrow L(D)$: Build initial classifier
2. While stopping-criterion is not satisfied do:
 - (a) $x \leftarrow S_L(X, D, h)$: Apply S_L , the selective sampling algorithm and get the next example from X
 - (b) $\omega \leftarrow f(x)$: Ask the oracle to label x
 - (c) $D \leftarrow D \cup \langle x, \omega \rangle$: Update the labeled example set
 - (d) $X \leftarrow X \setminus \{x\}$: Remove x from the unlabeled data
 - (e) $h \leftarrow L(D)$: Update the classifier
3. Return Classifier h

If this selection of samples for labeling is done well, a classifier with good generalisation performance can be built from a small training set. Thus the objective is to select examples that will be maximally informative for the classifier. The obvious way to do this is to select samples from the set X about which the classifier is least certain. When the classifier is a SVM this can be readily done by selecting the sample with the smallest classification margin. Research has shown that uncertainty in an ensemble of classifiers is also an effective criterion for sample selection in AL [9]. Alternatively, sample selection can be guided by a version space analysis where samples are selected that will most reduce the version space of the classifier [10].

In the work presented here we are interested in selecting a single informative example for a number of classifiers. Since we are primarily interested in the aggregation aspect we use a simple margin-based estimate of uncertainty (informativeness) for the component SVMs.

3.1 Multi-Label Active Learning with Support Vector Machines

The component classifiers used in this analysis are SVMs and for multi-label image classification a single SVM is trained for each of the k labels. In the evaluation presented in the next section we consider three sample selection strategies:

1. Random Selection: We randomly select an image from the unlabeled query pool and ask the user for feedback (i.e. the correct labels) and update the datasets i.e. adding the query image to the training set and removing it from the query pool set. The SVM models for each label are updated and this process is repeated for n steps, i.e. n requests to the oracle.
2. Single Label Annotation: This is the standard AL strategy where samples are selected to maximize informativeness for a single classifier. For each label independently, we calculate the SVM margin values for images in the query pool and select the one with lowest margin as query image. The training and query pool set are updated and the SVM models for each label are updated and this process is repeated for n steps.
3. Multi-label Annotation: We again select the query image based on the uncertainty where closeness to the SVM margin is the measure of uncertainty. But in this case, we calculate the SVM

margin values for all labels simultaneously and the distances from the margins are converted to a probability score. The probabilities across all SVMs are averaged to identify images that will be useful across all SVMs. The training and query pool set are updated and the SVM models for each label are updated. This process is repeated for n steps.

These three strategies are all reasonably well represented by the general AL process presented in Algorithm 1. However, the sample selection step 2(a) is different in each case. The strategy used for single label annotation is easily explained by reference to the standard SVM decision function:

$$h(x) = \text{sign} \left(\sum_{i \in SV} \alpha_i \omega_i K(x, x_i) \right) \quad (1)$$

In binary SVM classification the class labels are $\omega \in \{-1, +1\}$ and the class label is determined by the sign of the summation in equation 1 where SV is the set of support vectors, K is the kernel function and α_i is the learned weight for support vector i . The size of the summation gives the distance from the SVM margin so a policy of selecting the query example with the smallest margin will select samples on which the SVM is least certain.

For the multi-label annotation scenario we wish to aggregate these scores to select samples that will be informative for most classifiers. This is not straightforward as margin scores from different SVMs are not directly comparable. The strategy we use is to convert the SVM output to a posterior probability using a sigmoid function as proposed by Platt [11].

$$\Pr(y = 1|x) \approx P_{A,B}(f(x)) \equiv \frac{1}{1 + \exp(Af(x) + B)} \quad (2)$$

An improved strategy for learning the parameters A and B from the data is proposed by Lin *et al.* [12] – this is the strategy used in our implementation. By converting the SVM outputs to probabilities in this way we can calculate an aggregate uncertainty score for a sample across all labels by averaging these probabilities. This summarizes how the sample selection step, 2(a) in Algorithm 1, works in the multi-label situation. It is worth mentioning that the user query step 2(b) is also different in that the user returns a vector \mathbf{w} of labels, i.e. step 2(b) becomes $\mathbf{w} \leftarrow f'(x)$ where $f'(x)$ returns all labels for image x .

3.2 Related Work on Multi-Label Classification and AL

There are very few papers which address the problem of active learning in the context of multi-label classification. Boutell *et al.* [13] present a framework to handle systems where classes may overlap and apply it to scene (image) classification. The authors discuss approaches for training and testing in multi-label scene classification and introduce new metrics for evaluating individual examples, class recall and precision, and overall accuracy. They introduce *Cross Training*, a new training strategy to build classifiers, the *C-Criterion* for threshold selection using the MAP principle and *α -Evaluation*, a novel generic evaluation metric to evaluate multi-label classification results in a wide variety of settings. Elisseff and Weston [14] present SVMs but without AL for multiple labels based on a large margin ranking system. Li *et al.* [15] present query selection strategies for active learning for multi-label classification. Luo *et al.* [16] present an active learning method in multi-class classification problems to recognize underwater zooplankton from high-resolution images.

4 Evaluation

We tested our technique on the Barcelona image dataset as described in section 2.1 and Table 1. We represent each image by a feature vector of 286 dimensions using color, texture and edge features.

Specifically, we extract the following features from each image: color moments i.e. average, standard deviation, variance of RGB values and 128-dimensional global average RGB histogram; texture auto-correlation, texture co-occurrence matrix, texture edge frequency and 72-dimensional edge direction histogram using the Canny edge detector [17].

Our main objective in the evaluation is to be able to reduce the overall number of images that are presented to the user for labeling and show that active learning for multi-labels performs better than simple single-label active learning. We also use random selection as a baseline in the evaluation.

4.1 Experimental Setup

For each fold in the active learning evaluation, the dataset was randomly shuffled and partitioned as follows (see Table 1). A holdout test set of 27 images is kept back from the training process to assess generalization accuracy. The training data of 112 instances is divided into an initial labeled set of four images and an unlabeled query pool of 108 images.

Image Set	No. of Instances
Training Set	112
Test Set	27
Starting Pool Set	4
Query Pool Set	108

We use SVMs as a classifier as explained in section 3.1 because SVMs have shown to give better performance than other classifiers on similar problems [18]. We use our own implementation of the SMO algorithm for SVMs as presented by Platt [19]. The kernel employed in all evaluations is a Gaussian kernel with the kernel width set by cross validation.

In the following sections, we compare results with and without active learning on the Barcelona dataset and also compare results for single-label active learning and multi-label active learning. Specifically, we compare the base label accuracies with the number of query images required to achieve the baseline accuracy.

4.2 Without Active Learning

We first evaluate the Barcelona dataset using SVMs without active learning. Table 2 shows the 5-fold cross validation accuracy for each of base labels without active learning, which also acts as our baseline accuracy to be achieved with active learning. Results are shown in Section 4.5.1.

4.3 Single Label Annotation with Active Learning

For the single-label case (see section 3.1), we present a single query image, for each label independently, to the user at each iteration based on uncertainty where closeness to the SVM margin is the measure of uncertainty and ask the user for the correct image label and based on the given feedback, we rebuild our classifier with the new labeled image added to the training set. Figure 3 presents the accuracies for each label against the number of queries. The results for multi-label and random selection are also shown.

4.4 Multi-label Annotation with Active Learning

For the multi-label case (see section 3.1) we present a query image for feedback based on all the labels taken together, where distances from the margin are converted to a probability and the probabilities across all SVMs are averaged to identify images that will be useful across all SVMs. The results provided here are for the base-class evaluation. By base-class, we mean for each of the four labels in our dataset, we have four base classifiers and we calculate their individual accuracies on a test set. This evaluation strategy is taken from the work by Boutell *et al.* [13].

The evaluation strategy is as follows, let Y_x be the set of true labels for a test image x and P_x be the set of predicted labels from classifier h . Let Ω be the set of base-classes and ω is the output label for x . Let $H_x^\omega = 1$ if $\omega \in Y_x$ and $\omega \in P_x$, 0 otherwise. And let $\tilde{P}_x^\omega = 1$ if $\omega \in P_x$, 0 otherwise and likewise, let $\tilde{Y}_x^\omega = 1$ if $\omega \in Y_x$, 0 otherwise. Then [13], we calculate the following statistics for each label, on dataset T :

$$\text{Recall}(\omega) = \frac{\sum_{x \in T} H_x^\omega}{\sum_{x \in T} \tilde{Y}_x^\omega} \quad ; \quad \text{Precision}(\omega) = \frac{\sum_{x \in T} H_x^\omega}{\sum_{x \in T} \tilde{P}_x^\omega}$$

$$\text{Accuracy}_T = \frac{\sum_{x \in T} \sum_{\omega \in \Omega} H_x^\omega}{\max(\sum_{x \in T} \sum_{\omega \in \Omega} \tilde{P}_x^\omega, \sum_{x \in T} \sum_{\omega \in \Omega} \tilde{Y}_x^\omega)}$$

This evaluation measures the performance of the system based on the performance on each base label classifier.

4.5 Results

4.5.1 Without Active Learning:

Table 2 shows the cross validation baseline accuracies for each label. In section below, in some cases the active learning accuracy is higher than the base line accuracy and we present only averaged values for accuracies in the results for active learning.

Table 2: Cross Validation Accuracy without Active Learning for Labels

Target Label	Overlapping Examples	Features	Accuracy(%)
Buildings	108	286	86
Flora	65	286	65
People	61	286	60
Sky	117	286	98

4.5.2 With Active Learning for Single-Label and Multi-Label:

Figure 3 shows the accuracy for AL for each base label classifier against the number of queries performed. As shown in Figure 3(a) for label ‘Buildings’, multi-label annotation outperforms random selection. More importantly the number of image queries required to achieve the baseline accuracy is reduced significantly. For the ‘Buildings’ label, the baseline accuracy is reached around the 15th query and remains constant after that. As expected the single label accuracy reaches baseline accuracy even more quickly as the query image added during feedback steps is selected for each label independently (see section 3.1) and helps in optimizing the SVM model much more quickly. For label ‘Sky’ 3(d), multi-label annotation also performs much better than random given the number of queries and achieves baseline accuracy around 15th query image. On the other hand the multi-label active learning does not perform so well on the labels ‘People’ in Figure 3(c) and ‘Flora’ in Figure 3(b). It is probably not a coincidence that these are the labels on which the classifier has poorest performance

without any active learning. Our conclusion is that we need to improve baseline accuracy on these classes before active learning can work well and further more have better representing examples for the labels.

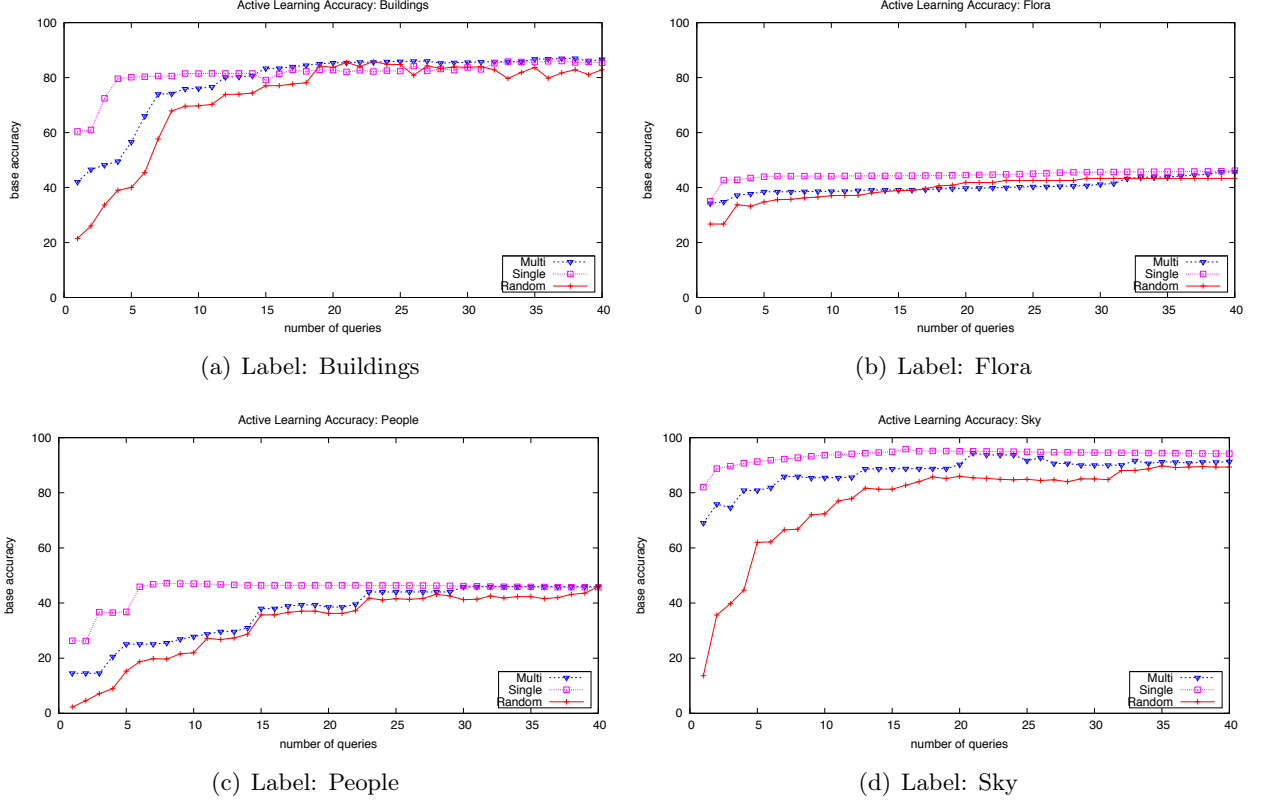


Figure 3: Active Learning: Base Label Accuracy vs the No. of Queries for Random Selection, Single Label Annotation and Multi-label Annotation (a), (b), (c), (d)

5 Conclusion and Future Work

In this paper, we have presented some preliminary results on using active learning for multi-label image annotation. Using SVMs as the base-classifier we have shown that the number of image queries to be presented to the user for labeling can be reduced significantly if a single query image is chosen for all classifiers together. We have shown that single label annotation provides better accuracy than multi-label annotation but multi-label annotation outperforms random selection. The main advantage of the proposed technique is that it successfully minimizes the overall number of images that the user is required to label in the training of the system. For these experiments we have chosen labels that are learnable from low-level color, texture and edge features. In the future we will work with labels that are more useful and more challenging and will extract features which help in representing the object categories present in the images more consistently using global and local image features and also better query selection techniques for active learning.

References

- [1] Kustanowitz, J., Shneiderman, B.: Motivating Annotation for Personal Digital Photo Libraries: Lowering Barriers While Raising Incentives. Univ. of Maryland Technical Report HCIL-2004 **18** (2005)
- [2] Matellanes, A., Evans, A., Erdal, B.: Creating an application for automatic annotation of images and videos. SWAMM (2006)
- [3] Datta, R., Ge, W., Li, J., Wang, J.Z.: Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In: MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia, New York, NY, USA, ACM Press (2006) 977–986
- [4] Chen, M.Y., Christel, M., Hauptmann, A., Wactlar, H.: Putting active learning into multimedia applications: dynamic definition and refinement of concept classifiers. In: MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, New York, NY, USA, ACM Press (2005) 902–911
- [5] Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture LIBRARIES. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(9) (2001) 947–963
- [6] Wu, Y., Kozintsev, I., Bouguet, J.Y., Dulong, C.: Sampling strategies for active learning in personal photo retrieval. In: Multimedia and Expo, 2006 IEEE International Conference on. (2006) 529–532
- [7] Sychay, G., Chang, E., Goh, K.: Effective image annotation via active learning. In: Multimedia and Expo. Volume 1. (2002) 209–212 vol.1
- [8] Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia, New York, NY, USA, ACM Press (2001) 107–118
- [9] Körner, C., Wrobel, S.: Multi-class ensemble-based active learning. In: ECML. ACM (2006) 687–694
- [10] Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: Proceedings of ICML-00, 17th International Conference on Machine Learning. Volume 2., Stanford, US, Morgan Kaufmann Publishers, San Francisco, US (2000) 45–66
- [11] Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers **10**(3) (1999) 61–74
- [12] Lin, H., Lin, C., Weng, R.: A note on Platt’s probabilistic outputs for support vector machines. Machine Learning **68**(3) (2007) 267–276
- [13] Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9) (September 2004) 1757–1771
- [14] Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in Neural Information Processing Systems. (2002) 681–687

- [15] Li, X., Wang, L., Sung, E.: Multilabel svm active learning for image classification. *Image Processing, 2004. ICIP '04. 2004 International Conference on* **4** (Oct. 2004)
- [16] Luo, T., Kramer, K., Goldgof, D.B., Hall, L.O., Samson, S., Remsen, A., Hopkins, T.: Active learning to recognize multiple types of plankton. *J. Mach. Learn. Res.* **6** (2005) 589–613
- [17] Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8**(6) (November 1986) 679–698
- [18] Wang, Y., Zhang, H.: Content-based image orientation detection with support vector machines. In: *Content-Based Access of Image and Video Libraries, 2001. (CBAIVL 2001). IEEE Workshop on.* (2001) 17–23
- [19] Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. MIT (1999) 185–208