Semi-supervised and Active Learning in Video Scene Classification from Statistical Features

Tomáš Šabata[1], Petr Pulc[2], and Martin Holeňa[2]

[1] Faculty of Information Technology, Czech Technical University in Prague,
Prague, Czech Republic
tomas.sabata@fit.cvut.cz

[2] Institute of Computer Science of the Czech Academy of Sciences,
Prague, Czech republic
{pulc,martin}@cs.cas.cz

Abstract. In multimedia classification, the background is usually considered an unwanted part of input data and is often modeled only to be removed in later processing. Contrary to that, we believe that a background model (i.e., the scene in which the picture or video shot is taken) should be included as an essential feature for both indexing and follow-up content processing. Information about image background, however, is not usually the main target in the labeling process and the number of annotated samples is very limited.

Therefore, we propose to use a combination of semi-supervised and active learning to improve the performance of our scene classifier, specifically a combination of self-training with uncertainty sampling. As a result, we utilize a combination of statistical features extractor, a feed-forward neural network and support vector machine classifier, which consistently achieves higher accuracy on less diverse data. With the proposed approach, we are currently able to achieve precision over 80% on a dataset trained on a single series of a popular TV show.

Keywords: video data, scene classification, semi-supervised learning, active learning, colour statistics, feedforward neural networks

1 Introduction

Automatic multimedia content labeling is still a comparatively difficult domain for machine learning. High input data dimensionality requires large training data sets, especially for approaches that are designed without prior assumptions on the data properties.

Moreover, the increasing resolution of image sensors brings higher detail (and thus, at least in theory, more information), but poses a significant issue for training phases of virtually all machine learning algorithms.

Many approaches, therefore, have to introduce a trade-off concerning the number of involved parameters, the number of distinct output labels (classes) [26] and the resolution of the input imagery [7]. Alternatively, they have to use only the statistical properties of the input data (as [3] and many others).

We also need to tackle the limitation on the amount of labeled training data.

Recent trends in video content processing include a task usually called Video to Text. The primary objective of such processing is to take multimedia content and describe its main features in a human-comprehensible text. Such representation may contain gathered information on the scene, actors, objects and actions in which they are involved. Such as the single image description "baseball player is throwing ball in game," as presented in [12].

Current approaches, however, commonly omit the information concerning the visual appearance of the background in complex multimedia content – even though such information might provide substantial contextual information for the object detection and event description itself. Approaches that use neural networks are mostly data-driven and require large amounts of data to adapt to each selected class. This requirement is, however, seldom met in smaller multimedia collections, such as home video, university lecture recordings, movie studios or corporate media databases.

We also want to reflect that a particular scene can be recalled by a human from a couple of static frames. Therefore, manual scene labeling is a relatively easy task as opposed to event labeling that may need the full video sequence or object labeling that commonly requires drawing a bounding box around the annotated object.

To use the limited human involvement in scene labeling as efficiently as possible, we employ semi-supervised learning to allow making use of unlabeled data, which are substantially easier to obtain, whereas simultaneously selecting the data for annotation using active learning methods.

The rest of this paper is organized as follows: In Section 2, we briefly summarize the state of the art in scene classification in the context of single images without significant obstruction by foreground objects, as well as the state of the art in combining semi-supervised learning (SL) and active learning (AL). Section 3 describes our approach to scene recognition in video content. In Section 4, we compare the accuracy of our method for different approaches to feature selection and different classifiers.

2 State of the Art

Scene recognition is rather simple from the human perspective. Whether the scene is the same as one previously visited is recognized by the overall layout of the space, presence, and distribution of distinct objects, their texture, and color. Other sensory organs can provide even more information and allow faster recall. Scenes not visited beforehand may fall after a thorough exploration into one of broader categories based on similarity of such features.

Multimedia content, however, does not allow such space exploration directly. It is constrained to the color information of individual pixels at a rather small resolution. Video content resolves this issue only partially with a motion of the camera, which, on the other hand, introduces more degrees of freedom in background modeling and increases its complexity.

2.1 Single Image Scene Classifiers Based on Colour Statistics

The early scene classifiers, including the Indoor/Outdoor problem [22,27], and also the more recent approaches mentioned below are directly based on the overall color information contained in the picture. The vital decision in this particular case is the selection of color space and the granularity of the considered histograms.

RGB (red, green and blue components) is the primary color space of multimedia acquisition and processing. However, it does not directly encode the quality of the color perceived by a human. By qualities of color, we primarily mean the color shade (hue). In HSV encoding (hue, saturation and value of the black/white range components, the last of them related to the overall lightness of the color), hue is commonly sampled with finer precision (narrower bins in histogram approaches) than saturation and lightness [5,8].

Mainly because of memory consumption and model size, statistical features of the individual images are commonly used for image processing, including basic scene classification. Other approaches are based on object detection [11,15], on interest point description [3,2], or in recent years they use deep convolutional neural networks [26,29,32].

2.2 Multi-label Extension

Often, a single image contains multiple semantic features – such as sea, beach and mountains. A crisp classification into only one class would, however, have to take only the dominant class, which might be different from the selection of the annotator. A somewhat possible extension is to create a new crisp class for each encountered combination of the labels, but this would have a substantial impact in the areas where the amount of labeled content is not sufficient to enable proper training on such sub-classes.

Another possibility is to organize the labels into a hierarchical structure. If the described scenery shares multiple features, the parent label may be preferred for content description. When the scene classifier detects only a specific part of the scenery, we should not consider it a full miss.

Statistical approach One of common assumptions in scene classification is that, during a single shot, the background will be visible for a more extended period than the foreground object. Therefore, we may process each frame in a single shot by a scene recognition algorithm and vote among the proposed labels. The statistical approach to background modeling applies if we assume a static camera shot. When such an assumption is met, all frames are perfectly aligned, and the background model can be extracted from the long-term pixel averages.

2.3 Semi-supervised Learning and Active Learning

Semi-supervised learning (cf. the survey [33]) is a technique that benefits from making use of easily obtainable unlabeled data for training. In this paper, we mainly focus on the self-training aproach to semi-supervised learning [10]. It is a simple and efficient method, in which we add samples with the most confidently predicted labels (pseudo-labels) to the training dataset. This can be done so the model is retrained in each iteration. Other aproaches to semi-supervised learning are co-training [1] and multiview training [9] thath benefit from agreement among multiple learners.

Active learning (cf. the survey [23]) is related to semi-supervised learning through being also used in machine learning problems where obtaining unlabeled data is cheap and manual labeling is expensive but possible. Its goal is to spend a given annotation budget only on the most informative instances of the unlabeled data. Most commonly, it is performed as *pool-based sampling* [14], assuming a small set of labeled data and a large set of unlabeled data. Samples that were found to be the most informative, are given to an annotator and are moved into the labeled set. The considered machine learning model (e.g., a classifier) is retrained and the algorithm iterates until the budget is exhausted or the performeance of the model is satisfactory.

Pool based sampling needs to evaluate an *utility function* that estimates some kind of usefulness of knowing the label of a particular sample. There are various ways of defining the utility function: for example, as a measure of uncertainty in *uncertainty sampling* [13], as a number of disagreements within an ensemble of diverse models in a method called *query-by-committee* [25], as the expected model change [24], the expected error [20] or only the variance part of the model error [6].

Semi-supervised and active learning can be quite naturally combined since they address unlabeled data set from opposite ends. For example, self-training uses the most certain samples to be turned to labeled samples and uncertainty sampling queries the most uncertain samples and obtains its label from an annotator. Such a combination was used for various problems [16,21,31]. Successful combinations with active learning exist also for multiview training [17,18,30].

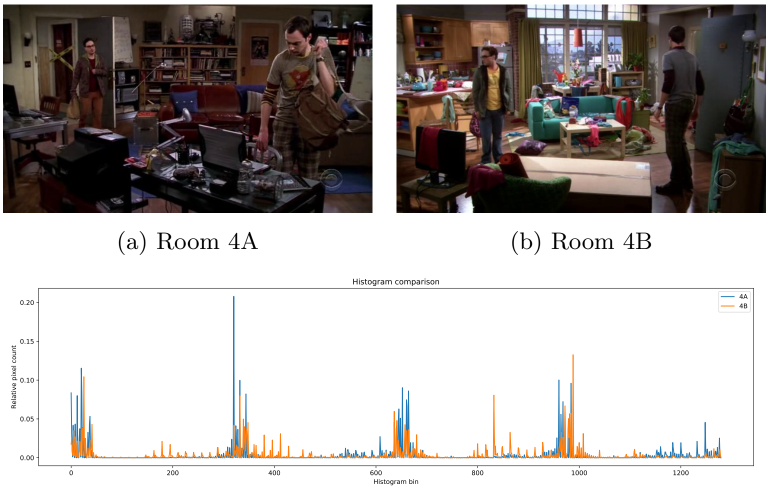
3 Multimedia Histogram Processing with Feed-Forward Neural Network using SVM

In the reported research, our main concern is to enable an automatic annotation of small datasets with a generally small variation within the individual classes. For example, we are not particularly interested in recognition of a broader scenery concept (such as a living room), but we aim at the classification that the video shot was captured in one specific living room.

One of the possible applications, on which we will demonstrate our approach in the next section, is the classification of individual scenes in long-running shows and sit-coms. However, our approach is designed to be versatile and enable, for example, disambiguation of individual television news studios or well-known sites.

Another concern of us is that the training of the classifier should require a minimal amount of resources to enable connection into more complex systems of multimedia content description as a simple high-level scene disambiguation module.

Therefore, we revise the traditional approaches in scene classification and propose the use of color histograms, possibly with partial spatial awareness. To demonstrate our reasoning behind this step, we refer to Figure 1.



(a) Room 4A

(b) Room 4B

(c) Histogram comparison

Fig. 1: Representative frames from two distinct living rooms and comparison of the proposed histograms. Although both of these pictures depict a living room, the distribution of colours is different. Source images courtesy of CBS Entertainment.

We choose a feed-forward neural network as the base classifier. In particular, we use a network with two hidden layers of 100 and 50 neurons and logistic sigmoid as activation function. The output layer uses the softmax activation function. The network is trained using backpropagation with a negative log-likelihood loss function and a stochastic gradient descent optimizer. The network topology, activation function and optimizer were found through a simple grid search, in which we considered also other the activation functions such as ReLU or hyperboilic tangent, and another optimizer, based on an adaptive sestimates of first and second moments of the gradients [**?**].

For the scene classification task, we can use the trained neural network directly. However, we introduce an improvement inspired by transfer learning. Transfer learning is usually used in deep convolution neural nets where the convergence of all parameters is slower [28]. However, we would like to demonstrate, that the transfer learning can bring a substantial benefit also in shallow neural networks. Especially in combination with a support vector machine (SVM) classifier.

In our scenario, we freeze the parameters of first layers and use the network as a feature extractor. For the classification stage, the original softmax layer is then replaced with a linear support vector machine. This brings us a rather small but consistent improvement in the final accuracy.

For an overall structure of our proposed network, please refer to Figure 2. In the figure, red arrows represent the first learning phase in which parameters of the net are found using a backpropagation. Blue arrows represent the second learning phase – transfer learning. In the second phase, the first two layers of the already trained neural net are used for training dataset generation. After that, a linear SVM classifier is trained. Green arrows represent the prediction of new samples.

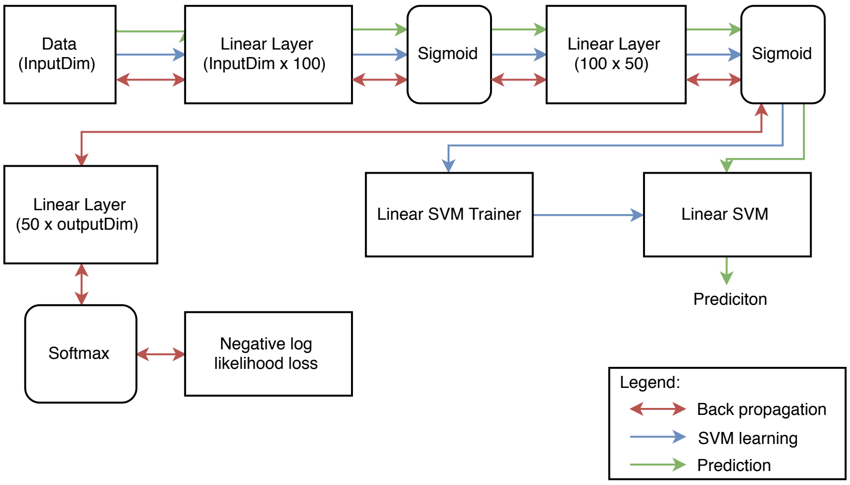


Fig. 2: The architecture of the proposed neural net. Red arrows represent the first learning phase; blue arrows represent a second learning phase with SVM and green arrows represent the prediction phase.

Finally, the model performance was improved by using a combination SL+AL. We have chosen a combination of uncertainty sampling with pseudo-labeling through self-training. In the experimental evaluation, the utility functions least uncertain (eq. 1), margin (eq. 2) and entropy (eq. 3) were included.

$$\phi_{LC}(x) = P_\theta(y_1^*|x), \tag{1}$$

$$\phi_M(x) = P_\theta(y_1^*|x) - P_\theta(y_2^*|x)), \tag{2}$$

$$\phi_E(x) = -\sum_i^N P_\theta(y_i|x)\mathrm{log}P_\theta(y_i|x), \tag{3}$$

In each iteration, n samples with the lowest utility function were queried to be annotated. At the same time, samples with the utility function higher than a threshold were predicted using the current version of the model, and these predictions were then used to train the next version of the model. Utility functions were calculated from the output of softmax layer of the neural net. The number of samples n was chosen to be 5 in each iteration. The threshold value was tuned to keep the number of wrong labels getting into training data as low as possible.

3.1 Weighted accuracy

The scene description in our experiment is constructed hierarchically so there are three different levels of the label. The first level describes building name, the second level describes a room, and the last level describes detail in the room. For instance, if the camera shot captures the whole living room of the flat “4A” in the “main” building, we use a label such as `main.4a`. If only a specific portion of the room is shown, we use a more detail level of the label such as `main.4a.couch`.

To take into account the label hierarchy, we introduce weighted accuracy of a classifier F predicting $\hat{y}_1, \ldots, \hat{y}_n$ for training data $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\mathrm{WA}(F) = \frac{1}{n}\sum_{i=1}^{n} f(y_i, \hat{y}_i),$$

$$f(y_i, \hat{y}_i) = \begin{cases} 1 & \text{if } \mathbb{1}(y_i = \hat{y}_i, 3) \\ 0.5 & \text{if } \mathbb{1}(y_i = \hat{y}_i, 2) \\ 0 & \text{otherwise.} \end{cases},$$

where $\mathbb{1}(y_i = \hat{y}_i, k)$ is the truth function of equality of all components of y_i and $\hat{y}_i$ on the k-th or a higher level of the component hierarchy.

4 Experimental evaluation

For the evaluation of all the following approaches, we prepared our dataset [19] from the first series of a sit-com The Big Bang Theory. This particular show uses only a couple of scenes and by 2018 new series are still being produced. The dataset is chosen for the proof of concept experiment and new datasets should follow in future experiments. The multimedia content was automatically segmented into individual camera shots by PySceneDetect [4] using the content detector.

A middle frame from the detected shot was stored as a reference for human annotation and convolutional neural network processing. Due to the copyright protection, these stored frames are not contained in the dataset. They were divided into 80% training and 20% test data along the time axis.

For statistical approach experiments, the following histograms averaged by the respective frame area and shot duration were obtained: *RGB 8x8x8* (flattened histogram over 8 × 8 × 8 bins), *H* (hue histogram with 180 bins), *HSV* (concatenation of 180 bins H, 256 bins S and 256 bins V histograms) and _HSV 20x4x4 2*2_ (flattened histogram over 20 × 4 × 4 bins in each of 4 parts of the frame introduced by its prior division in 2 × 2 grid).

4.1 Combinations of histograms and classifiers

We have compared combinations of the above described histograms with the following classifiers: linear SVM, k nearest neigbours (k-NN), naive Bayes (NB) and the feedforward neural nets (FNNs) described in section 3, i.e., FNN alone and FNN+SVM. A full comparison of the unweighted accuracy of all 16 combinations is carried out in Table 1.

Table 1: Accuracy of combining the considered four kinds of histograms with the following classifiers: linear SVM, k-NN, NB, FNN and FNN+SVM. For each classifier, the highest accuracy with respect to the different kinds of histograms is in italics, and the highest accuracy with respect to different classifiers is in bold

Accuracy [%]	Linear SVM	k-NN	NB	FNN	FNN+SVM
RGB 8x8x8	18.1	32.42	26.1	54.5	**60.0**
H	12.4	30.7	26.1	56.0	**58.9**
HSV	14.1	32.6	32.4	63.3	**65.7**
HSV 20x4x4 2*2	*46.0*	*45.4*	*33.9*	*77.2*	***78.8***

It is noticeable that HSV 20x4x4 2*2 feature dominates over all other variants. Therefore, we were using HSV 20x4x4 2*2 in the subsequent experiments. On the other hand, adding an SVM as the last layer of the FNN brings only a smaller improvement.

4.2 Comparison with an inception style neural network

State-of-the-art approaches in image scene classification usually use the residual deep convolutional neural networks with inception-style layers. They are typically combinded with multi-scale processing of the input imagery.

With these key features in mind, we used the winner of the 2016 LSUN challenge [29] as the reference method for scene classification on our dataset.

The results are, however, worse than expected. The accuracy progress (see Figure 3) shows that the network training is very unstable. The testing accuracy achieves a maximum of 32.4% in the 801st epoch.

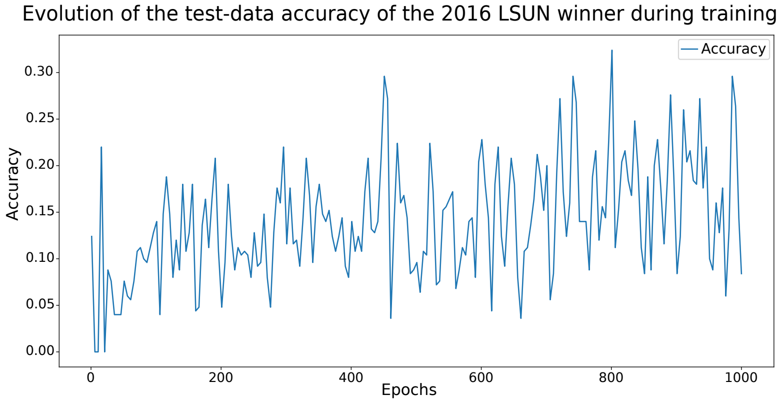


Fig. 3: Accuracy of the inception-style winner of the LSUN challenge [29] on the testing set

As we are unable to interpret the inner state of the neural network directly, we may only assume that the main issue with using the multi-resolution convolutional neural network is the small dataset size. However, this is exactly the issue we need to mitigate.

4.3 Including supervised and active learning

As was shown in Subsection 4.1, the use of feed-forward neural network itself brings a substantial increase in classification metrics. As Table 2 indicates, the SVM layer provides an additional improvement as well as using part of the unlabeled dataset with SL+AL. Although the improvement is not high, we believe that using the more sophisticated combination of SL+AL could bring us even further.

The initial labeled dataset contained 5315 samples. An unlabeled dataset with 26528 samples was used for both active and semi-supervised learning. A human annotator was asked five queries at each of ten iterations.

Table 2: Final achieved accuracy, weighted accuracy, precision, Recall and F1 score with the HSV 20x4x4 2*2 histogram. For each of these classifier performance measures, the highest value among the considered classifiers is in bold

	Acc	Weighted acc	Precision	Recall	F1
FNN	0.7723	0.8518	0.7813	0.7590	0.7578
FNN-SVM	0.7883	**0.8626**	0.8026	0.7837	0.7842
FNN-SVM with SL+AL	**0.7895**	0.8617	**0.8037**	**0.8022**	**0.7978**

5 Conclusions and Future Work

In this paper, we sketched how semi-supervised learning combined with active learning can be applied to scene recognition In addition, we propose to use neural networks for further feature enhancement.

The resulting features extracted from the proposed neural network provide a substantial improvement over the engineered features on input. Especially, if the extracted features are used as a data embedding for a linear SVM classifier.

This allows us to achieve an accuracy of almost 79% on a small dataset that is significantly higher than reference method (32.4%).

Several descriptors are, however, still hard to recognize even for a human annotator (e.g. staircase floor number). In these situations, one may benefit from the context of the previous and following shot and consequently improve the classification accuracy. Therefore, we would like to try context-based classifiers, such as HMM, CRF or BI-LSTM-CRF as a next step of our research.

Last but not least, we would like to use transductive SVM in the top layer of the final classifier and provide further experiments in the combination with semi-supervised and active learning, primarily with active multiview training.

Acknowledgements

The reported research has been supported by the grant 18-18080S of the Czech Science Foundation (GAČR).

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100. ACM (1998)
2. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via plsa. In: European conference on computer vision. pp. 517–530. Springer (2006)
3. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE transactions on pattern analysis and machine intelligence **30**(4), 712–727 (2008)
4. Castellano, B.: Pyscenedetect. `https://github.com/Breakthrough/PySceneDetect` (2017)

5. Chen, L.H., Lai, Y.C., Liao, H.Y.M.: Movie scene segmentation using background information. Pattern Recognition **41**(3), 1056–1065 (2008)
6. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning **15**(2), 201–221 (1994)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
8. Fan, J., Elmagarmid, A.K., Zhu, X., Aref, W.G., Wu, L.: Classview: hierarchical video shot classification, indexing, and accessing. IEEE Transactions on Multimedia **6**(1), 70–86 (2004)
9. Farquhar, J., Hardoon, D., Meng, H., Shawe-taylor, J.S., Szedmak, S.: Two view learning: Svm-2k, theory and practice. In: Advances in neural information processing systems. pp. 355–362 (2006)
10. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in neural information processing systems. pp. 529–536 (2005)
11. Han, S., Kim, J.: Video scene change detection using convolution neural network. In: Proceedings of the 2017 International Conference on Information Technology. pp. 116–119. ACM (2017)
12. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
13. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine Learning Proceedings 1994, pp. 148–156. Elsevier (1994)
14. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 3–12. Springer-Verlag New York, Inc. (1994)
15. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Advances in neural information processing systems. pp. 1378–1386 (2010)
16. Liu, A., Jun, G., Ghosh, J.: A self-training approach to cost sensitive uncertainty sampling. Machine Learning **76**(2), 257–270 (Sep 2009). https://doi.org/10.1007/s10994-009-5131-9, `https://doi.org/10.1007/s10994-009-5131-9`
17. Mao, C.H., Lee, H.M., Parikh, D., Chen, T., Huang, S.Y.: Semi-supervised co-training and active learning based approach for multi-view intrusion detection. In: Proceedings of the 2009 ACM Symposium on Applied Computing. pp. 2042–2048. SAC '09, ACM, New York, NY, USA (2009). https://doi.org/10.1145/1529282.1529735, `http://doi.acm.org/10.1145/1529282.1529735`
18. Muslea, I., Minton, S., Knoblock, C.A.: Active + semi-supervised learning = robust multi-view learning. In: Proceedings of the Nineteenth International Conference on Machine Learning. pp. 435–442. ICML '02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002), `http://dl.acm.org/citation.cfm?id=645531.655845`
19. Pulc, P.: Replication data for: Feed-forward neural networks for video scene classification from statistical features (2018). https://doi.org/10.7910/DVN/MPZGWO, `https://doi.org/10.7910/DVN/MPZGWO`
20. Roy, N., McCallum, A.: Toward optimal active learning through monte carlo estimation of error reduction. ICML, Williamstown pp. 441–448 (2001)

21. Sabata, T., Borovicka, T., Holena, M.: K-best viterbi semi-supervized active learning in sequence labelling (2017)
22. Serrano, N., Savakis, A., Luo, A.: A computationally efficient approach to indoor/outdoor scene classification. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on. vol. 4, pp. 146–149. IEEE (2002)
23. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1), 1–114 (2012)
24. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in neural information processing systems. pp. 1289–1296 (2008)
25. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 287–294. COLT '92, ACM, New York, NY, USA (1992). https://doi.org/10.1145/130385.130417, `http://doi.acm.org/10.1145/130385.130417`
26. Song, F.Y.Y.Z.S., Xiao, A.S.J.: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
27. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on. pp. 42–51. IEEE (1998)
28. Tang, Y.: Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239 (2013)
29. Wang, L., Guo, S., Huang, W., Xiong, Y., Qiao, Y.: Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. IEEE Transactions on Image Processing **26**(4), 2055–2068 (2017)
30. Wang, W., Zhou, Z.H.: On multi-view active learning and the combination with semi-supervised learning. In: Proceedings of the 25th international conference on Machine learning. pp. 1152–1159. ACM (2008)
31. Yao, L., Sun, C., Wang, X., Wang, X.: Combining self learning and active learning for chinese named entity recognition. Journal of software **5**(5), 530–537 (2010)
32. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
33. Zhu, X.: Semi-supervised learning literature survey (2005)