Face Recognition via Heuristic Deep Active Learning

Ya Li[1], Keze Wang[2], Lin Nie[2], and Qing Wang[2(✉)]

[1] Guangzhou University, Guangzhou 510006, China
liya@gzhu.edu.cn
[2] Sun Yat-sen University, Guangzhou 510006, China
wangkeze@alumni.sysu.edu.cn, nie.lin@foxmail.com,
wangq79@mail.sysu.edu.cn

Abstract. Recent successes on face recognition tasks require a large number of annotated samples for training models. However, the sample-labeling process is slow and expensive. An effective approach to reduce the annotation effort is active learning (AL). However, the traditional AL methods are limited by the hand-craft features and the small-scale datasets. In this paper, we propose a novel deep active learning framework combining the optimal feature representation of deep convolutional neural network (CNN) and labeling-cost saving of AL, which jointly learns feature and recognition model from unlabeled samples with minimal annotation cost. The model is initialized by a relative small number of labeled samples, and strengthened gradually by adding much more complementary samples for retraining in a progressive way. Our method takes both high-uncertainty samples and the high-confidence samples into consideration for the stability of model. Specifically, the high-confidence samples are selected in a self-paced learning way, and they are double verified by the prior knowledge for more reliable. These high-confidence samples are labeled by estimated class directly, and our framework jointly learns features and recognition model by combining AL with deep CNN, so we name our approach as heuristic deep active learning (HDAL). We apply HDAL on face recognition task, it achieves our goal of "minimizing the annotation cost while avoiding the performance degradation", and the experimental results on Cross-Age Celebrity Dataset (CACD) show that the HDAL outperforms other state-of-the-art approaches in both recognition accuracy and annotation cost.

Keywords: Active learning · Deep CNN · Self-paced learning · Face recognition

1 Introduction

With the growth of mobile phones and social networks, it is fairly easy to obtain facial images. The demands of developing intelligent systems with face recognition technology are increasing accordingly. Traditional approaches handle this problem by supervised learning. However, it is a slow and expensive process to label these images

J. Zhou et al. (Eds.): CCBR 2017, LNCS 10568, pp. 97–107, 2017.
https://doi.org/10.1007/978-3-319-69923-3_11

for preparing a good labeled training dataset. An effective approach to reduce the annotation effort is active learning (AL), which helps learner to select the most informative samples and obtains a high recognition performance.

The basic principle of AL methods is to progressively select and annotate the most informative unlabeled samples, and boost the current model by them in an incremental way. The sample selection criteria are extremely important in AL. Specifically, the high-uncertainty samples, together with other criteria like density and diversity of categories distribution under current model, are generally treated as the informative candidates for model retraining. However, the existing AL approaches all neglect the role of high-certainty samples. The recently proposed self-paced learning (SPL) algorithm [1, 2] demonstrates the important role of high-certainty samples. The SPL presents the training data in an easy to difficult order [3, 4], which imitates the learning process of humans. An easy sample in SPL is the one with high prediction confidence by current model. It is interesting that the AL and SPL actually select samples in the opposite criteria. We want to investigate the possibility of making them complementary to each other.

On the other hand, the existing AL approaches are limited by hand-craft features and small-scale datasets. As well known, the learned features and joint architectures of deep learning methods have made dramatic progress on many vision tasks, especially the deep convolutional neural network (CNN) methods. But a deep CNN model requires many more labeled samples than shallow structure. Those learned features by CNN are updated all the time with the classifier's upgrading, and traditional AL methods can't provide sufficient samples for CNN fine-tuning and make it difficult to obtain the optimal feature representation. Using a batch of high-confidence samples in a self-paced way can bridge the gap of deep CNN and AL well. We think that these high-confidence samples play an important role in the stability of the model, and adding them into training set will reduce the annotation effort further. However, it will result in the deviation problem if assigning the estimated labels as supervised information directly because of the low reliability of the initial model. Therefore, we must take certain measures to ensure the prediction accuracy.

In this paper, we propose a useful framework combining the deep CNN and AL in a self-paced fashion. The framework jointly learns features and recognition model from unlabeled samples with minimal annotation cost. Unlike the existing AL methods only selecting the high uncertainty samples, our method also takes the high-confidence samples as complementary samples for better stability and robustness of model. We employ the dynamic confidence threshold on the sample selecting stage. With the model's performance improving, the samples selection threshold decreases correspondingly. Specifically, considering that the initial model is unreliable and tends to deviate by outliers, we take the prior knowledge into consideration. The high-confidence samples are further ranked by the distance to labeled samples. More close to samples of same identity, more reliable. Those samples both with high-confidence and high distance rank are labeled by predicted category, we called the labels as pseudo labels, and we name our approach as heuristic deep active learning (HDAL). By using softmax output as category probability, the AL method can be easily combined with deep CNN. Our HDAL approach handles both manually annotated and pseudo-labeled samples simultaneously.

There are two contributions of this work: (1) we propose a useful framework combining the AL with CNN, which makes it is possible in the large-scale scenarios by using pseudo-labeled samples for the upgrading of model; (2) we applies the novel framework into face recognition task, and the experiments on Cross-Age Celebrity Dataset (CACD) [5] show that our approach outperforms other methods in both recognition accuracy and annotation cost.

2 Related Work

The samples selection criteria are extremely important for AL. One of the most common strategies is the uncertainty sampling. Lewis et al. [6] selected uncertainty samples by category probability with probabilistic classifier. A more general method is using entropy to realize uncertainty sampling. Joshi et al. pointed out in multi-class cases, the entropy values were heavily influenced by probability values of unimportant classes, and they proposed a Best-versus-Second-Best (BvSB) approach to address this problem in [7], which took the difference between top two high estimated probability values as uncertainty measure. There are many methods using Query-by-committee (QBC) [8, 9], which select samples those have high classification variance. SVM-margin based method [10] took the samples closer to decision boundary as high uncertainty samples.

Many works incorporate density of unlabeled samples into AL [11–13]. Settles and Craven [11] weighted an unlabeled sample by its average similarity to other unlabeled samples. Compared with cosine similarity used in [11], the work [12] used mutual information density and the work [13] used clustering-based density information.

Moreover, there are some researches [13–15] take the samples diversity into consideration for class balance. Brinker [14] considered the angles between the induced classification hyperplanes, where each newly chosen sample corresponded to a hyperplane which maximizes the minimum angle to previous hyperplanes. Elhamifar et al. [15] captured the distribution of samples with low confidence scores. Demir and Bruzzone [13] selected the samples at center of K-means clusters for diversity.

The methods mentioned above all ignore the "from easy to complex" learning process of human. The inspiration of SPL can be explained in analogous to human cognitive process. Bengio et al. [1] initialized the conceptual learning paradigm as curriculum learning (CL), the key in which is to find a ranking function that assigns learning priorities to training samples. Kumar et al. [2] designed a new formulation for adjusting the predetermined curriculum by the feedback about the learner, named SPL. Jiang et al. formulated SPL as a concise optimization problem [3], and further discovered the missing link between CL and SPL [4].

3 Our Approach

In this section, we illustrate how our HDAL model works. Suppose we have a face image set containing n images of m persons, the label of image x_i is person ID j, that is, $y_i = j, j \in \{1, \ldots, m\}$. Let the labeled sample set is $\mathcal{L}$, unlabeled sample set is $\mathcal{U}$, and

the current classification model is $\mathcal{M}$. The HDAL for face recognition is formulated as follows:

$$\min_{\theta} -\frac{1}{n}[\sum_{i=1}^{n}\sum_{j=1}^{m} 1\{y_i = j\}p(y_i = j \mid x_i; \theta)], \tag{1}$$

where θ is the network parameters of the CNN, $1\{\cdot\}$ is indicator function, and $p(y_i = j \mid x_i; \theta)$ is the softmax output of CNN, which represents the probability of estimating x_i as j class.

3.1 CNN Classification Model

We use CNN classification model for retraining, which contains 8 layers, the front 5 layers are convolution-pooling layers, next 2 layers are fully-connected layers and the last one is softmax output layer. Figure 1 shows the overall network architecture. Neurons in two fully-connected layers are dropped out by 50%.

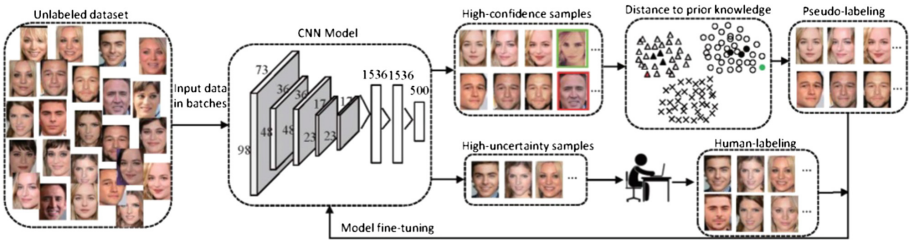


Fig. 1. Illustration of our propose HDAL framework.

Our HDAL has 3 main stages: (1) Initialization. For each class we randomly select a small number of unlabeled samples and manually annotate them as training samples to initialize the CNN. (2) Complementary samples selection. We first rank all unlabeled samples by the current CNN model: for high-confidence samples, we further calculate their distances to prior knowledge and keep samples with low distance, then assign their pseudo-labels directly using estimated labels; for high-uncertainty samples, we deliver them to human annotator. (3) CNN fine-tuning. All labeled complementary samples are put into CNN for retraining. The complementary samples selecting and CNN fine-tuning are executed alternatively until the model converged.

3.2 Sample Selection Criteria

High-Uncertainty Sample Selection

We extend the common uncertainty sampling criteria to select top-K uncertainty samples instead of only one in an iteration considering the convergence rate of large-scale database. We introduce three common active learning criteria, which are based on the probability of a sample class.

(1) Least confidence (LC). It selects the top-K samples, which are ranked by the most likely class probability in descending order. Suppose y^* is the most likely class label for sample x^*. LC can be formulated as:

$$x^* = \underset{x \in U, \text{topK}}{\text{argmax}}(1 - p(y^* \mid x; M)), \tag{2}$$

(2) Best vs second-best (BvSB). Suppose y_1^* and y_2^* are the first and the second likely class label of x^*. The smaller difference of top two classes probabilities, the much more uncertain the category.

$$x^* = \underset{x \in U, \text{topK}}{\text{argmax}} -(p(y_1^* \mid x; M) - p(y_2^* \mid x; M)). \tag{3}$$

(3) Entropy measure (EN). Entropy is often used as an uncertainty measure in informatics. The larger sample entropy, the much more uncertain the category.

$$x^* = \underset{x \in U, \text{topK}}{\text{argmax}} - \sum_i p(y_i^* \mid x; M) log p(y_i^* \mid x; M). \tag{4}$$

High-Confidence Sample Selection

The high-confidence samples with their pseudo-labels are used for retraining directly. To ensure the reliability of pseudo-labels, high-confidence and distance to prior knowledge are taken into account jointly. We employ a dynamic confidence threshold δ, which is decreased in decay rate β with iteration step t. Feeding the training data in an easy to complex order, it imitates the learning process of human like SPL. Denote the high-confidence candidates set as $\mathcal{H}'$, it is formulated as:

$$\begin{aligned} \mathcal{H}' &= \{x^* \mid p(y^* \mid x; M) > \delta\} \\ \delta &= \delta - \beta t. \end{aligned} \tag{5}$$

Suppose the predicted label of a selected sample x^* is p, p=1, 2, …, m. Then we compute the distance between x^* and samples centroid of same identity P, and the distance between x^* and samples centroid of different identities $\bar{P}$. The lower difference of above two distances means that the more reliable of the predicted label. We rank the difference values of all the selected samples and use them for double verification. We further select top-k high-confidence samples $\mathcal{H}$ and assign them the pseudo-labels. The distance to prior knowledge criterion can be formulated as:

$$\mathcal{H} = \{x^* \mid \underset{x \in H', \text{topk}}{\text{argmin}}[(x_p^* - P) - (x_p^* - \bar{P})]\}, \tag{6}$$

The recognition model can be retrained after adding these new labeled samples, which include the high-uncertainty samples U_c annotated by human annotator and the high-confidence samples $\mathcal{H}$ with pseudo-labels. The whole HDAL algorithm is shown in Algorithm 1.

Algorithm 1 The Whole HDAL Algorithm

Input: labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$, initialized model $\mathcal{M}$

Output: the face recognition model $\mathcal{M}$

1: for $t = 1$ to T do /* T is the maximum iteration */

2: $\mathcal{M} = \text{train}(\mathcal{L}, \mathcal{M})$; /* model fine-tuning, see Algorithm 2 for detail */

3: $P = \text{test}(\mathcal{U}, \mathcal{M})$; /* using $\mathcal{M}$ to estimate the class probability of $\mathcal{U}$ */

4: $[U_c, \mathcal{H}] = \text{select}(\mathcal{U}, P, \text{strategy})$; /* select high-uncertainty samples U_c by Eq.(2)-Eq.(4) and select high-confidence samples $\mathcal{H}$ by Eq.(5)-Eq.(6) */

5: query(U_c); /* resort to annotator */

6: pseudoLabel($\mathcal{H}$); /* assign the pseudo-labels */

7: $\mathcal{L} = \mathcal{L} + H + U_c$; /* update $\mathcal{L}$ */

8: end for

3.3 Parameter Optimization

Adding the complementary samples into the labeled sample set $\mathcal{L}$, we fine-tuning the CNN model iteratively. Suppose the number of samples in $\mathcal{L}$ is increased to N, the cost function of our HDAL is rewritten as follows:

$$J(\theta) = -\frac{1}{N}[\sum_{i=1}^{N}\sum_{j=1}^{m} 1\{y_i^* = j\} p(y_i^* = j \mid x_i^*; \theta)], \tag{7}$$

where $p(y_i^* = j \mid x_i^*; \theta) = \frac{e^{\theta_j^T x_i^*}}{\sum_{l=1}^{m} e^{\theta_l^T x_i^*}}$.

There is no closed-form solution for θ, we therefore resort to gradient descent algorithm and employ the standard back propagation to update the CNN's parameters θ. The partial derivative of the network parameters θ is:

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{N}\sum_{i=1}^{N} x_i^*(1\{y_i^* = j\} - p(y_i^* = j \mid x_i^*; \theta)). \tag{8}$$

Then update θ_j by Eq. (9) on each iteration.

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} \quad j = 1, 2, \ldots, m. \tag{9}$$

The fine-tuning of CNN model is realized in Algorithm 2.

Algorithm 2 CNN Model Fine-tuning Algorithm

Input:

labeled training set X, initial parameters θ, and learning rate α

Output:

network parameters θ

1: for $s = 1$ to S do /* S is the maximum iteration */

2: $P = \text{test}(X, \theta)$; /* estimate the class probability of X */

3: $G = -\frac{1}{N} \times X \times \left(1 - P^s\right)$; /*compute gradient by Eq.(8) */

4: $\theta = \theta - \alpha G$; /* update parameters by Eq.(9) */

5: end for

4 Experiments

We present experimental results for the proposed HDAL on face recognition task using CACD[1] database [5]. The CACD database is a large-scale database released in 2014, which contains more than 160,000 images of 2,000 celebrities. There are only 200 celebrities' images are manually checked originally, and we extend the number to 580. Among them, the images of 80 individuals are utilized for pre-training of feature presentation, and the rest 500 persons' images are used to perform the HDAL approach. The 10% images with their labels are used to initialize the CNN model, and the rest 90% images are performed a five-folder cross-validation.

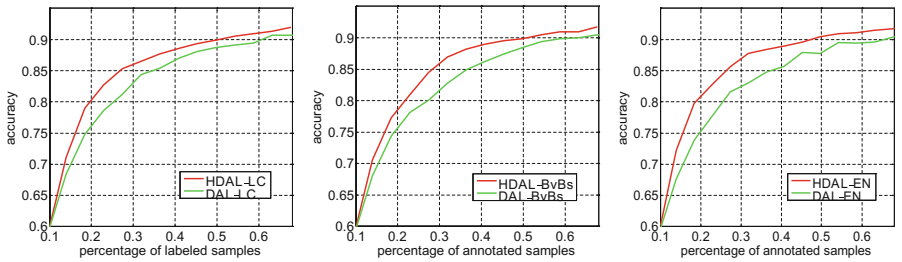
Our CNN model is constructed based on Caffe [16], the initial parameters are set by Gaussian distribution $\mathcal{N}(0, 0.01)$, and the learning rates of all the layers are set as 0.01. The experiments are executed on a PC with Nvidia Titan X GPU. We first detect the faces using the method proposed in paper [17] and resize the faces to 150×200. In each iteration, the number of high-uncertainty samples K is set as 1000; the selection threshold of high-confidence samples δ are set as 0.98, and the reduced rate of threshold β is set as 0.0033.

First a baseline experiment is conducted, which train the CNN model with 80% labeled images and test the rest 20% images. The recognition rate of baseline method is 92%, which can be considered as the best performance of CNN model can reach.

Then we verify the effectiveness of the high-confidence samples selection criterion by a set of experiments. According to different high-uncertainty samples selection

[1] http://bcsiriuschen.github.io/CARC/.

strategies, our approach further is named as HDAL_LC, HDAL_BvSB and HDAL_EN. For the traditional AL method without using the high-confidence samples, we name them as DAL_LC, DAL_BvSB and DAL_EN. Figure 2 illustrates the performance comparison between our heuristic selection strategy and traditional active learning strategies. The subfigures (a)–(c) demonstrate different high-uncertainty sampling criteria: (a) is LC, (b) is BvBS, and (c) is EN. To achieve 85% recognition accuracy, the labeled training samples required for HDAL_LC, HDAL_BvBS, HDAL_EN are 28.7%, 28.2% and 26.8% respectively, while for DAL_LC, DAL_BvBS, DAL_EN are 34.4%, 36% and 37%. When the 50% training samples are labeled, the recognition accuracies of HDAL_LC, HDAL_BvBS, HDAL_EN are 90%, 90% and 90.5% respectively, while for DAL_LC, DAL_BvBS, DAL_EN are 88.8%, 88.4% and 87.9%. To reach the baseline recognition accuracy 92%, HDAL_LC, HDAL_BvBS and HDAL_EN require 67%, 67.6% and 68% labeled training samples respectively. We can see that the performance of our HDAL is much better than the DAL.



(a) LC (b) BvBS (c) EN

Fig. 2. The comparison of recognition accuracies between HDAL and DAL.

We further evaluate the contributions comes from different components of HDAL. First using SPL and AL standalone respectively and then using their combination. For AL method, we tried the random and entropy-based selection strategies, which are denoted as RAND and DAL. RAND points to randomly selecting the samples to be labeled by annotator. Figure 3 illustrates the accuracies obtained by using SPL, AL and HDAL. The accuracies of SPL, RAND, DAL and HDAL are 71.4%, 78.2%, 82.5%, and 86.9% respectively, when the percentage of labeled samples is 30%. AL can resort annotator for the informative samples labeling, so RAND and DAL are both better than SPL. RAND and DAL gain 6.8% and 11.1% accuracy improvement over SPL. The combination of SPL and DAL, that is HDAL, can automatically exploit the majority of the high-confidence samples, and further achieves 4.4% accuracy improvement over DAL.

At last we compare our HDAL with other active learning methods, such as TCAL [13], CPAL [15], and RAND. CPAL annotates samples in each step based on prediction uncertainty and sample diversity. And TCAL takes uncertainty, diversity and density into account jointly; it outperforms other state-of-the-art methods. We

re-implement CPAL and TCAL by using our CNN model without last softmax output layer as feature representation for fair comparison. RAND method is randomly selecting the samples to be labeled for CNN fine-tuning. The performance of RAND can be regarded as the worst performance in active learning methods. The comparison results are shown in Fig. 4. Our HDAL model outperforms the competing methods in accuracy when the same amount labeled samples. When the 50% training samples are labeled, the recognition accuracy of TCAL, CPAL and RAND are 87.5%, 87.9% and 84.9% respectively. On the other hand, our HDAL reduces the annotation effort compared to other method. To achieve 85% recognition accuracy, TCAL, CPAL and RAND requires 39.5%, 40.8% and 49.8% labeled samples respectively, while HDAL only requires 26.8%. Our HDAL shows the better performance than the TCAL and CPAL on both accuracy and labeling cost.

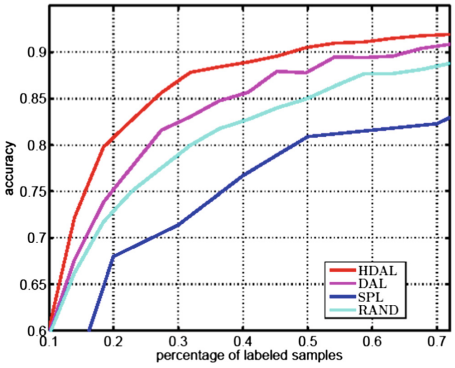


Fig. 3. Accuracies of different components with labeled samples increasing.

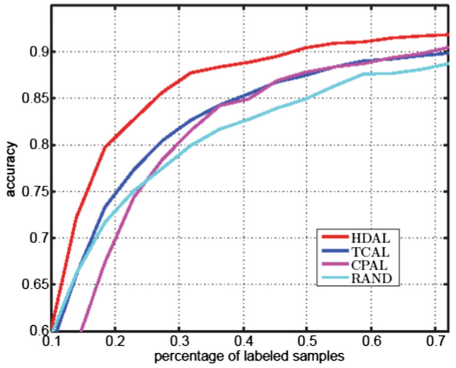


Fig. 4. The comparison with other state-of-the-art methods.

From above experiments, one can see that our HDAL is better than other methods in both recognition accuracy and annotation cost. We think that the good performance of HDAL stems from the better discrimination by CNN model and the better robustness by jointly considering high-confidence samples and high-uncertainty samples.

5 Conclusion

In this paper, we propose a novel heuristic deep active learning approach and apply it in face recognition task. Our HDAL framework combining AL with deep CNN jointly learns features and recognition model from unlabeled samples with minimal annotation cost. We take the high-confidence samples as complementary samples for better stability and robustness of model compared with the traditional AL methods, which only take the high-uncertainty samples into consideration. Specifically, the high-confidence samples are selected in a self-paced learning way, and they are double verified by the

prior knowledge for more reliable. By using softmax output as category probability, HDAL combines the deep CNN with AL successfully. The better discrimination of CNN model and the better robustness of jointly considering high-confidence samples and high-uncertainty samples make it outperform other state-of-the-art methods. In future, we plan to apply our HDAL approach on more challenging and general object recognition task. We also plan to generalize our framework into other vision tasks.

Acknowledgments. This research is supported by the Research Project of Guangzhou Municipal Universities (No. 1201620302), National Undergraduate Scientific and Technological Innovation Project (No. 201711078017), the Science and Technology Planning Project of Guangdong Province (Nos. 2015B010128009, 2013B010406005). The authors would like to thank the reviewers for their comments and suggestions.

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML, pp. 41–48. ACM (2009)
2. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: NIPS, pp. 1189–1197 (2010)
3. Jiang, L., Meng, D., Mitamura, T., Hauptmann, A.G.: Easy samples first: self-paced reranking for zero-example multimedia search. In: MM, pp. 547–556. ACM (2014)
4. Jiang, L., Meng, D., Zhao, Q., Shan, S., Hauptmann, A.G.: Selfpaced curriculum learning. In: AAAI, pp. 2694–2700. AAAI Press (2015)
5. Chen, B.-C., Chen, C.-S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 768–783. Springer, Cham (2014). doi:10.1007/978-3-319-10599-4_49
6. Lewis, D.D.: A sequential algorithm for training text classifiers: corrigendum and additional data. In: ACM SIGIR Forum, pp. 13–19. ACM (1995)
7. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: CVPR, pp. 2372–2379. IEEE (2009)
8. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Mach. Learn. 133–168 (1997)
9. McCallumzy, A.K., Nigamy, K.: Employing em and pool-based active learning for text classification. In: ICML, pp. 359–367. Citeseer (1998)
10. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 45–66 (2002)
11. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Empirical Methods in Natural Language Processing, pp. 1070–1079. Association for Computational Linguistics (2008)
12. Li, X., Guo, Y.: Adaptive active learning for image classification. In: CVPR, pp. 859–866. IEEE (2013)
13. Demir, B., Bruzzone, L.: A novel active learning method in relevance feedback for content-based remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 2323–2334 (2015)
14. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: ICML, pp. 59–66 (2003)

15. Elhamifar, E., Sapiro, G., Yang, A., Sasrty, S.S.: A convex optimization framework for active learning. In: ICCV, pp. 209–216 (2013)
16. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: MM, pp. 675–678. ACM (2014)
17. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR, pp. 3476–3483. IEEE (2013)