Common Class Information based Efficient Training Data Selection for Photo Categorization

Ryo Shirasu
Nagoya Univercity
Chikusa-Ku, Nagoya, Japan
shirasu@mv.ss.is.nagoya-u.ac.jp

Jien Kato
Nagoya Univercity
Chikusa-Ku, Nagoya, Japan
jien@is.nagoya-u.ac.jp

Yu Wang
Nagoya Univercity
Chikusa-Ku, Nagoya, Japan
ywang@mv.ss.is.nagoya-u.ac.jp

Kenji Mase
Nagoya Univercity
Chikusa-Ku, Nagoya, Japan
mase@nagoya-u.jp

Abstract

In this paper, we proposed a novel method to efficiently select discriminative training samples for local photo classification or management. We introduced a concept widely shared by most of images: underlying common classes, and based on them we catch more colorful and more characteristic/discriminative training samples. We conduct multi-clustering with feature selection and adaptive sampling to the images of each single common class, and then adapt acquired information/knowledge to target local photos. We have evaluated proposed method on 21,424 photos taken in daily life of a nursery school for two classification problems. Experimental results show that our method is superior to the case without consideration of common classes or simply selecting samples randomly.

1 Introduction

Recent years, with the rapid spread of digital cameras, the amount of photos individuals accumulate has increased sharply. This situation leads to pressing needs for effective methods or technologies to efficiently manage and organize photos. Image labeling, the basis of the image classification and retrieval, is a promising solution for such kinds of the demands. However, since the semantic labels different people want to use in photo management will be very different, and thus the amount of the training data will get too large to be prepared manually, image labeling becomes a significantly severe problem.

Simple semantic tags are easily labeled by associating images together based on the similarity of contents. For example, Google Picasa [1] is a practical application of this kind of approaches, which utilizes face detection and user input tags to efficiently manage photos by suggesting the tags of annotated photos to the visually similar non-annotated ones. On the other hand, because in practical use, the semantic tags people actually want to annotate to personal photos are much more ambiguous and complicated, it is not easy to annotate these tags to unlabeled photos only based on simple similarity of image contents. To figure out this problem, photo classification with user predefined classes and a labeled training dataset will be a useful solution. However, taking this approach will causes two conflicting requirements: 1) it is desirable to label as many as possible training samples for good performance of classifiers, and 2) it is desirable to label as little as possible training samples for reducing user's load.

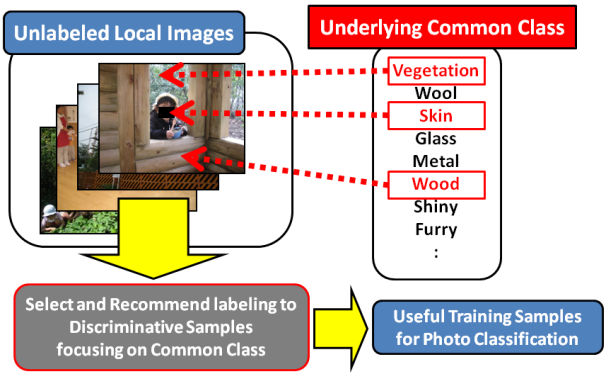

Figure 1. Overview of our proposed method.

There are two major methodologies, semi-supervised learning [2] and active learning [3], which can break up this dilemma and build relatively good classifiers using less labeled training data. In the former method, not only labeled but also unlabeled samples are used to train the classifiers. For instance, Self-Training algorithm [2] first predicts the label of unlabeled samples, and then adds the samples with higher prediction confidence to the training dataset to boost the classifier. On the other hand, in the latter method, useful samples that most likely contribute to the performance of the classifier are found out from unlabeled samples, and then are suggested to be labeled by users. Both of above approaches are classifier-based learning method that can obtain a good classifier using a large amount of unlabeled samples, from a relatively poor classifier trained by using a small amount of labeled samples.

With the same goal of coping with increasing classifier's performance and reducing users' load both, this paper proposes a novel training data selection method to select promising samples that most likely contribute to the classification of local images such as photo album. So, it is able to work jointly with all the learning methods that require labeled training data, including above-mentioned semi-supervised learning and active learning.

Figure 1 shows the overview of proposed method. Our method introduces a concept widely shared by most of images: underlying common classes, and we select characteristic/discriminative training samples focusing on these classes. For example, most of character photos seem to have the presence of the common class "human skin", and most of swimming pool photos seem to have the presence of the common class "water". To make full use of these common classes,

in our approach, we first extract two kinds of information/knowledge from underlying common classes of labeled images, that is, 1) discriminative feature dimensions, and 2) characteristic sampling methods in which representative samples can be chosen. Then, we adapt the information/knowledge to target local images to select promising training samples.

Comparing with semi-supervised learning and active learning that solve the problem like how to boost the target classifiers, since our proposed method does not need any information about target labels, but select promising samples from the target image set based on their various visual aspects, it can provide good training samples (recommend to be labeled by users) to initialize the classifiers for these existing learning methods.

The rest of the paper is organized as follows. In Section 2, the core of this paper, parameter learning and discriminative sample selecting are described. Experimental results and evaluation are discussed in Section 3. Finally, we draw our conclusion and future work in Section 4.

2 Proposed Method

Given totally unlabeled target local images such as photo album as input, proposed method selects promising samples, which most likely contribute to the classification of local images according to predefined classes by users, and recommend these samples to be labeled. In photo classification, the samples that are discriminative in visual aspects could be useful training data. So, in our proposed method, we seek discriminative samples by clustering target local images, and provide them to users. To find out such samples, we first conduct multi-clustering with feature selection and adaptive sampling in proper underlying common classes, and then utilize the information acquired to find out discriminative samples from target local images. That leads to the advantage of obtaining more various discriminative and characteristic samples than directly applying the clustering algorithm to the target local images.

Our proposed method consists of two phases: learning from common classes that is described in Section 2.1, and multi-clustering and adaptive sampling that is described in Section 2.2. In the phase of learning from common classes, we learn to select discriminative samples from the viewpoint of common classes, through evaluating classification performance on some public dataset with the same labels as the common classes. Although there are no labels, target local images also have underlying common classes. Therefore, in the phase of multi-clustering and adaptive sampling, we select and recommend useful training samples using the information/knowledge learned in the previous phase. Throughout the whole process, we use Repeated Bisection algorithm [9], one of partitional clustering algorithms, for clustering. It is capable of measuring cosine similarity between samples, and automatically deciding the number of clusters by similarity between clusters.

2.1 Learning from Common Classes

In this phase, we learn to select discriminative training samples from the viewpoint of individual common classes, through evaluating a labeled public dataset with the same labels as the common classes (we call the dataset "common class images"). Figure 2 shows the outline of learning. Here, there are two main factors we want to learn, both of them decide how well selected samples can contribute to the classification of target local images. We describe them as follows.

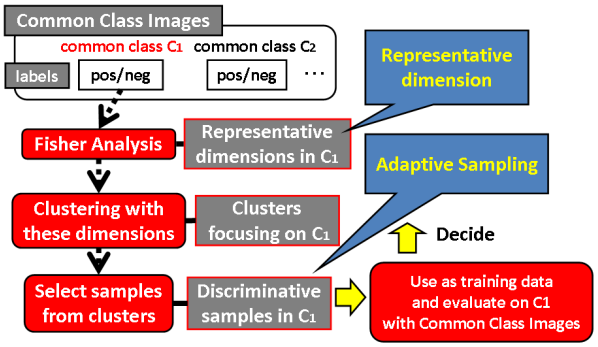

Figure 2. Overview of learning from common classes.

1) Dimensions to be used in clustering: d_t^*

In a feature vector calculated from the data related to a common class c_t, there usually exist some representative and some non-representative dimensions for representing this class. We believe that the representative dimensions for common classes will also contribute to finding characteristic/discriminative samples of target images by clustering them focusing on these common classes. Let d_t^* indicate the optimal dimensions to represent common class c_t and contribute to its discrimination. In order to find clusters in target images from the viewpoint of underlying common classes $\{ c_1 \ldots c_T \}$, we conduct clustering on a public datasets with the same labels as the common classes (i.e., common class images) to find $\{ d_1^* \ldots d_T^* \}$ by feature selection.

2) Adaptive sampling method from clustering results: s_t^*

To choose representative samples from clustering results, there are mainly two kinds of sampling methods: choose the samples near to centroids or choose those far from centroids. Additionally, the variety of training data may affect classification performance because of possible existence of overfitting. Similar to dimension selection, we believe that it is rational to select samples from the target local images in the way that is efficient to select representative samples in the underlying common classes, and it is better to learn the sampling method from data. Let s_t^* indicate how to select samples adaptively with respect to class c_t, from the viewpoint of the distance between samples and centroids. We learn the optimal sampling method s_t^* from common class images after clustering the data in dimensions of $\{ d_1^* \ldots d_T^* \}$.

The processing in phase of learning from common classes runs as below. For each c_t, proposed method seeks promising samples by varying the number of samples n at a certain number of dimension d and a certain sampling method s_t. The dimension number d and sampling method s_t are evaluated by the classification performance conducted to all the common class images, and the optimal one is chosen. Concretely, for each common class c_t, we firstly apply Fisher Analysis to the common class images, and select d dimensions of features (indicated by d_t) that make intra-class variance minimal and inter-class variance maximal. Sec-

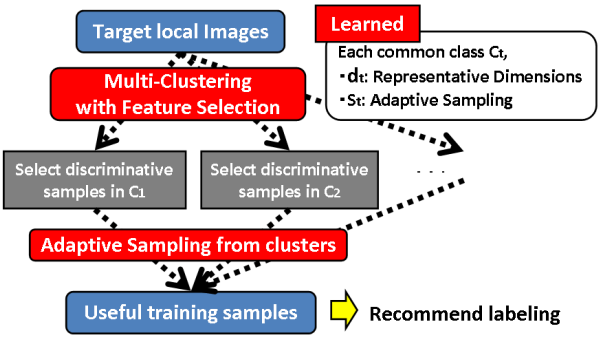

Figure 3. Overview of multi-clustering and adaptive sampling process.

ondly, we apply clustering to the same images by only using selected dimensions d_t and obtain the clusters focusing on class c_t. After that, based on current s_t, we select n samples from these clusters and evaluate the performance on all the common class images. Here, we preserve the combination of d_t and s_t with the best evaluation score for current n. Repeating this process, we finally find the best d_t and s_t as d_t^* and s_t^* for all possible n.

As a result, the optimal $\{ d_1^* \dots d_T^* \}$ and $\{ s_1^* \dots s_T^* \}$ that represent information/knowledge of common classes are learned for all the classes $\{ c_1 \dots c_T \}$.

2.2 Multi-Clustering and Adaptive Sampling

In this phase, given unlabeled target local images as input, proposed method selects useful training samples focusing on each of $\{ c_1 \dots c_T \}$ by feature selection based multi-clustering and adaptive sampling. Figure 3 shows the overview of this processing. In Section 2.1, proposed method has already learned two factors to select discriminative samples focusing on $\{ c_1 \dots c_T \}$. Here, we describe how to utilize $\{ d_1^* \dots d_T^* \}$ to find the clusters by multi-clustering in target local images, focusing on common classes $\{ c_1 \dots c_T \}$, and how to utilize $\{ s_1^* \dots s_T^* \}$ to adaptively select useful samples from the results of clustering, with respect to each of common classes $\{ c_1 \dots c_T \}$.

The processing in phase of multi-clustering and adaptive sampling runs as below. First, for each c_t in $\{ c_1 \dots c_T \}$, we apply feature selection to target images based on d_t^*. Secondly, we conduct multi-clustering individually in target images using feature vectors after feature selection by d_t^*. And then, we select equal number of training samples according to $\{ s_1^* \dots s_T^* \}$ from the clustering results focusing on each class in $\{ c_1 \dots c_T \}$ respectively. Finally, selected training samples are recommended to users to be labeled for further photo classification into desirable classes.

3 Experimental Results

In this section, we evaluate our proposed method on the photo album taken in daily life of a nursery school.

3.1 Experimental Settings

In this experiment, we try to select promising training samples from a photo album of a nursery school that consists of 21,424 images, and then use them to evaluate classification performance in the way the nursery school teachers want to use for photo management. The details are described as follows.

Target Local Images:

We collect 21,424 photos that reflect the daily life of a nursery school as the target local images. The photos include the scenes like playing, dancing, lunch, or some events such as Halloween, PE festival, graduation ceremony, and so on. Additionally, except children and teachers, parents, animals and other exterior people will rarely appear on the photos. We divide these photos into half, one for the input of proposed method for sample selection, and the other for evaluation. Each includes 10,712 images.

Common Classes:

We adopt PASCALVOC2008 [4] together with the annotations reported in [5] as common class images. In [5], the authors focused on the "attributes" of objects in the images and annotated labels to each bounding box of the objects. There are 6,340 objects labeled by three kinds of attributes: shape, part and material. From the three attributes, we think material will be useful and pick out 3 materials, "Skin", "Vegetation" and "Wood", as the common classes for target photos. And, totally 6,340 images belonging to 13 common classes are used in the experiments.

In phase of learning from common class, for each common class c_t, we vary the number of dimension d from 100 to 500, and vary the sampling method s_t within the following three choices: 1) select nearest samples to centroids, 2) select farthest samples from centroids, and 3) select both of nearest and farthest samples by half. In order to decide $\{ d_1^* \dots d_T^* \}$ and $\{ s_1^* \dots s_T^* \}$, we evaluate the classification performance by varying the sample number n from 100 to 1000 per 100, and decide both by majority vote.

Feature Vector:

For one image, we calculate following four kinds of features and concatenate them into a 1440-dimensional vector.

1) Color Histogram:
Composed of 512 bins in RGB channels.

2) Edge:
16-dimensional vector extracted by Canny edge detector on 4 by 4 regions.

3) SIFT(bag-of-visualwords):
Apply 128-dimensional SIFT descriptor [7] at every single keypoint in an image, and then quantize them into a histogram with 400 bins [6].

4) GIST:
512-dimensional-vector extracted by Gabor filter with various directions and frequencies [8].

Photo Classification:

We train SVM using a Radial Basis Function as kernel with training samples selected by proposed method using whole feature. By this SVM, we classify photos in a photo album of a nursery school into the classes defined by two ways:

1) Class definition I:
if a photo is group picture (includes more than four people) or not.

2) Class definition II:
if a photo is taken indoor or outdoor.

These ways for photo management are actually wanted by nursery school teachers and we know that through an interview with them. Given test images, the classification performance on those classes definition is evaluated by Area Under Curve (AUC) of SVM.

Baselines:

We compare our proposed method with following two baseline methods.

1) No Feature Selection:

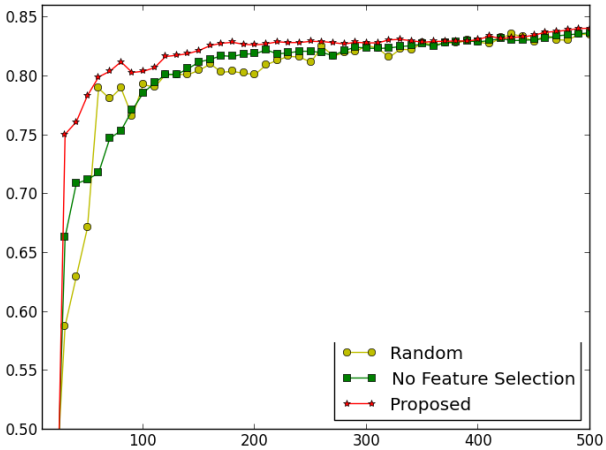

Figure 4. Evaluation results with class I.

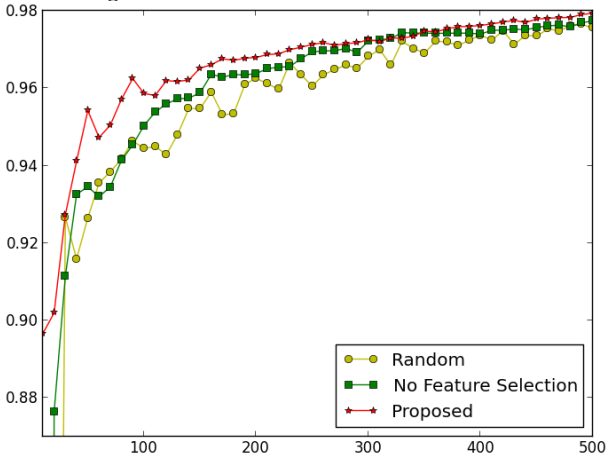

Figure 5. Evaluation results with class II.

without consideration of common classes, simply apply clustering to the whole feature vector, and select centroids as training data.

2) Random:

without searching promising training data, simply select samples randomly five times, and get their mean AUC.

3.2 Results and Discussion

Figure 4 shows AUC of SVM using our proposed method and baseline methods for class definition I, and Figure 5 for class definition II, increasing the number of training samples from 10 to 500 per 10 images.

In both situations, proposed method outperforms baselines especially at the small number of training samples up to 100. That exactly agrees with our goal since it means proposed method can achieve good performance of classification with less human labeling efforts.

Comparing with random method that does not search useful training samples, it is also clear proposed method certainly chooses samples that contribute to the classification. Further, the performance of proposed method is much more stable than random method as expected.

Comparing with no feature selection method that has not introduced the idea of common classes, proposed method conspicuously benefits from the information/knowledge of common classes in the meaning of selecting useful samples. On the other hand, we can find some less stable places in the AUC curves by proposed method. We think this comes from the complexity of sample selection, namely, since proposed method integrates a number of clustering results, some bias could occur in sample selection. We want to deal with this problem by introducing more proper common class in the future.

Through experiments with various combinations of common classes, we found the classification performance benefits a lot by introducing the common classes that frequently appearing in the target images. For instance, by visual observation, we have confirmed the nursery school photos used in this experiment include many appearances of leafs and plants that likely correspond to the common class "Vegetation", many appearances of persons that likely correspond to the common class "Skin", and also many appearances of floor/table that likely correspond to the common class "Wood". How to select optimal common classes will become a future subject.

4 Conclusion

In this paper, we proposed a novel method to efficiently select discriminative training samples for local photo classification or management. For this purpose, we introduced a concept widely shared by most of images: underlying common classes, and based on them we can catch more colorful and more characteristic/discriminative training samples. We conduct multi-clustering with feature selection and adaptive sampling to the images of each single common class, and then adapt acquired information/knowledge to target local photos. Experimental results show that our proposed method is superior to both of no feature selection method and random sample selection method.

As the future work, we will figure out the problem about how to decide common classes. For practical use, a way to associate specific target images with their underlying common classes is required. And furthermore, it is necessary to evaluate our method on more classification problems to verify its generalization ability.

Acknowledgement

This research was supported by the National Institute of Information and Communication Technology (NICT).

References

[1] Google Picasa: `http://picasa.google.com`
[2] X.Zhu: Semi-Supervised Learning Tutorial, ICML2007
[3] S.Dasgupta, et al.: A tutorial on active learning, ICML2009
[4] Visual Object Classes Challenge 2008: `http://pascallin.ecs.soton.ac.uk/`
[5] A.Farhadi, et al.: Describing Objects by their Attributes, Computer Vision and Pattern Recognition, pp.1778-1785, 2009.
[6] E.Nowak, et al.: Sampling Strategies for Bag-of-Features Image Classification, ECCV 2006.
[7] DG.Lowe: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision archive Volume 60 Issue 2, November 2004 Pages 91 - 110
[8] A.Oliva, et al.: Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision, Vol. 42(3): 145-175, 2001.
[9] Y.Ziao, et al.: Comparison of agglomerative and partitional document clustering algorithms, Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2002.