Multi-class Co-training learning for object and scene Recognition

(Xian-Hua Han)
College of Information Science and Engineering,Ritsumeikan University, 525-8577, JAPAN
Email:hanxhua@fc.ritsumei.ac.jp

(Yen-Wei Chen)
College of Information Science and Engineering,Ritsumeikan University, 525-8577, JAPAN

(Xiang Ruan)
Omron Cooperation, Japan

Abstract

It is often tedious and expensive to label large training data sets for learning-based object and scene recognition systems. This problem could be alleviated by semi-supervised learning techniques, which can automatically select more training samples from unlabel data for reducing the cost of labeling. In this paper, we proposed a multi-class co-training learning method of two different views for improving the performance of selective training samples for object and scene classification. In the co-training procedure, the classifiers are learned in two different views, respectively, and then, are used for classifying the unlabel data. At the same time, according to the confidence factor of the classified unlabel samples, we can confirm if the classifiers of the two views are enough strong for co-training or which is more stronger for co-training. Therefore, the unlabeled samples, which are classified by the strong classifier, can be chosen to label. To evaluate the performance of the proposed co-training method, two dataset (one is scene dataset, the other is object dataset) are utilized for recognition. The experimental results demonstrated that the recognition rate can be improved by co-training learning in different views, and it is also comparable with those by the art of the state algorithms.

1 Introduction

Image category recognition is important to access visual information on the level of objects (motorbikes, cars, etc.) and scene-like types (beaches, mountains, foods, etc.), and it has a wide range of applications, such as intelligent image processing and content-based image indexing and retrieval (CBIR) [1]. In CBIR, an efficient and effective classification method can significantly improve the retrieval accuracy by removing the irrelevant images. Image classification has posed a significant challenge to the research community of computer vision due to interclass variability, illumination and scale changes. The rich context of an image makes the semantic-understanding(scene and object recognition) very difficult. In the other hand, most developed and popular algorithms for image recognition are learning based model, in which a lot of pre-labeled training samples are needed for obtaining efficient classification model. However, it is often tedious and expensive to label large training data sets for learning-based method. This problem could be alleviated by semi-supervised learning techniques[2,3], which can automatically select more training samples from unlabel data for reducing the cost of labeling.

Usually, the feature representation of an image is a combination of diverse features, such as color, texture, shape, which represent global information of images [2]. Recently obtaining local descriptor of images become very hot research topic for image representation, in which the bag-of-word model [3] is the most success one for computer vision application. For a specified example, the contribution of different features is significantly different. So the training model from different feature is also different in recognizing the image properties. In this paper, we proposed a multi-class co-training learning method of two different views for improving the performance of selective training samples for object and scene classification. In learning procedure, global feature, such as color and edge, and local feature such as Bag-of-word model can be naturally considered as sufficient and uncorrelated views of an image. So a multi-category classification models are learned in global and local feature subspaces, respectively, and then, are used for classifying the unlabel data. At the same time, according to the confidence factor of the classified unlabel samples, we can confirm if the classifiers of the two views are enough strong for co-training or which is more stronger for all of the multi-class images. Therefore, the unlabeled samples, which are correctly classified by the strong classifier, can be chosen to label. In order to deal with the unbalance problem of selected training sample, we simultaneously select the fix number of training samples for each class in each iteration for updating. To evaluate the performance of the proposed co-training method, two dataset (one is scene dataset, the other is object dataset) are utilized for recognition. The experimental results demonstrated that the recognition rate can be improved by co-training learning in different views, and it is also comparable with those by the art of the state algorithms.

2 Feature extraction for image multi-view representation

In this section we describe how we extract feature for multi-view representation, which represent different visual properties of images, and at the same time, include enough discriminative information for image category. As we known, it is difficult to classify image recognition only with one type of image feature. So in this paper, we represent images with differen types of images features: popular local feature: bag-of-words model; global features such as color histogram and edge histogram, and merge them together for recognition.

(1) Bag-of-words feature: In computer vision, local

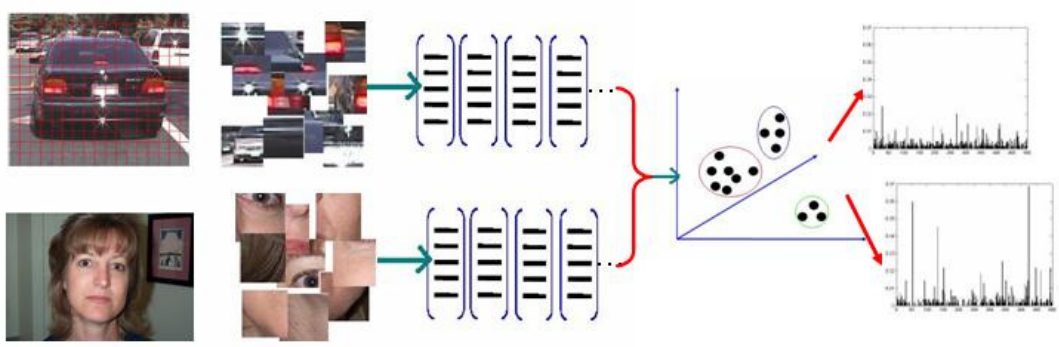

Figure 1. BOW feature representation. First column: original images and grids sampling for patches; Second column: Sampling patches; Third column: SIFT descriptor; Forth column: K-means cluster; Fifth column: bag-of-words histogram representations.

descriptors (i.e. features computed over limited spatial support) have proved well-adapted to matching and recognition tasks, as they are robust to partial visibility and clutter. Bag-of-words model in computer vision is the image representation of local descriptor histogram for object recognition and image classification. Figure 2 shows the procedure bag-of-words(BoW) feature extraction and the extracted histogram feature of example images [4].

(2) Color histograms are widely used to capture the color information in an image. They are easy to compute and tend to be robust against small changes of camera viewpoints. Given an image $\mathbf{I}$ in some color space (e.g., red, green, blue). The color channels are quantized into a coarser space with k bins for red, m bins for green and l bins for blue. Therefore the color histogram is a vector $\mathbf{h} = (h_1, h_2, \cdots, h_n)^T$, where $n = kml$, and each element h_i represents the number of pixels of the discretized color in the image.

(3) Edge histogram in segmented regions: We first segment the image into grid region, and then calculate edge histogram in each region for image representation. With the distribution over edge orientations within a region, Local shape can be captured for image representation, and the contacted shape features of all region are as the final image property for learning classification model.

3 Multi-class Co-training method

Co-training is a semi-supervised learning technique that requires more than two views of the data. It assumes that each example is described using at less two different feature sets that provide different, complementary information about the instance. Ideally, the two views are conditionally independent (i.e., the two feature sets of each instance are conditionally independent given the class) and each view is sufficient (i.e., the class of an instance can be accurately predicted from each view alone)[4]. For image representation, we can consider the two types features(global–color or edge histogram, and local–Bag-of-model feature) can provide different, complementary information for images, and It is natural and reasonable to assume that they are uncorrelated views of an image. the Co-training first learns a separate classifier for each view using any labeled examples. The most confidence predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data. However, the conventional co-training method only deal with two-class classification problem. It is not suitable for the multi-class object and scene recognition problem. So in this paper, we extend the two-class co-training method for multi-class recognition problem.

Assume that $\mathbf{x} = \{g_1, \cdots, g_i, c_1, \cdots, c_j\}$ is the feature representation of an image, where $\{g_1, \cdot, g_i\}$ and $\{c_1, \cdot, c_j\}$ are global attributes and local attributes of an image, respectively. For simplicity, we define the feature representation space $\mathbf{V} = \mathbf{V}_G \times \mathbf{V}_C$, and $\{g_1, \cdots, g_i\} \in \mathbf{V}_G$, $\{c_1, \cdots, c_j\} \in \mathbf{V}_C$.

In order to find high confidence-degree images as training sample from unlabel data, multi-class SVM is used to learn a classifier h_k on these labeled samples in the two feature subspace, respectively ($k = 1, 2$, the classifiers of two-view feature). The unlabeled set can be denoted by the probabilities belonging to each category of images, then the top N high probabilities P_k representing high confidence-degree samples for each category are used for confirming if the classifier h_k is a strong classification model.

(1) If all element of P_k are larger than a fixed threshold θ, we consider that the classification model with the k^{th} view is enough strong for the image recognition, and then, we update the $N \times M$ samples from the unlabel set as the training samples, where N is the updated sample number of each category from each classifier, M is the category number.

(2) If some element of P_k are smaller than the fixed threshold θ, we consider that the classification model with the k^{th} view is not enough strong for the image recognition,we do not select samples as training using the high probability data from the classifier h_k.

(3) If all classification model can not satisfy the requirement in (1), we will select samples from unlabel data using the high-confidence samples in the comparable strong view classifier, where the element number of P_k larger than the threshold θ is more.

It can be seen that the training samples for each category are updated in the same number in each iteration. Therefore the unbalance training sample problem is well dealed with our proposed multi-class co-training method. The labeling of training sample by the strong classification model will result in more strong and stable classifier for specific application. The global dia-

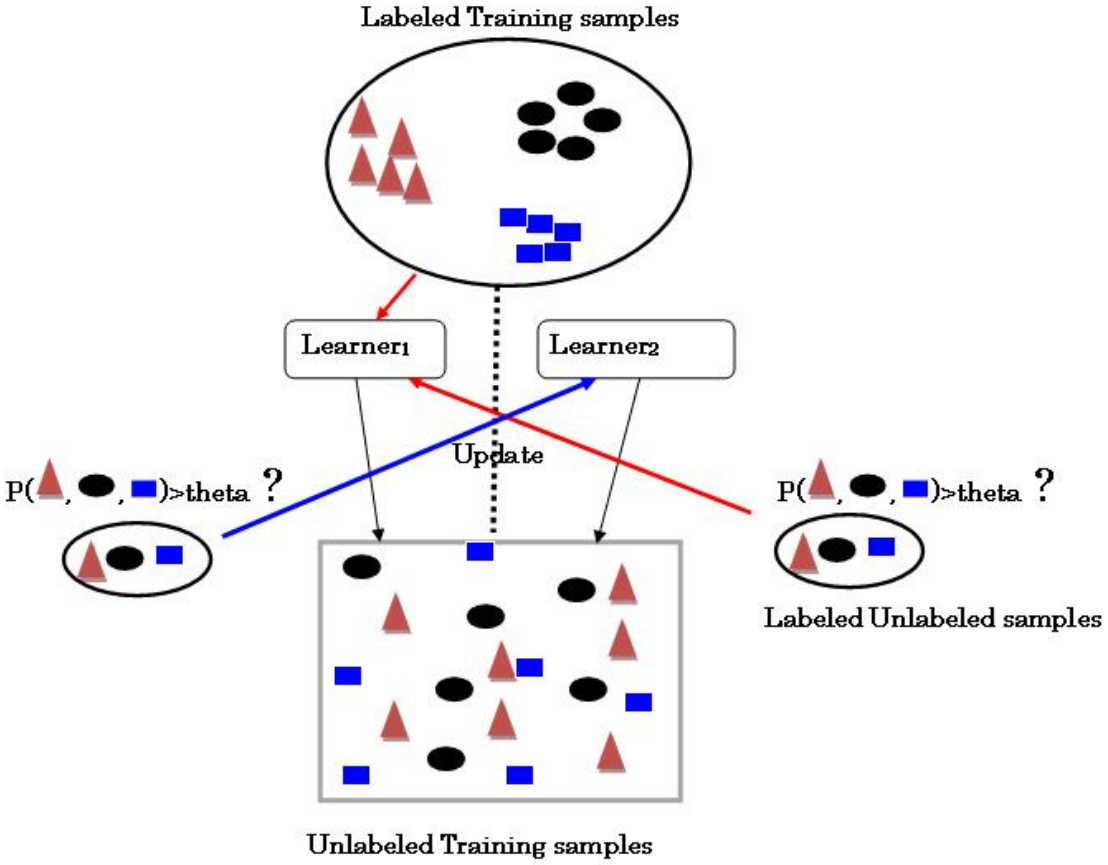

Figure 2. Multi-class co training procedure

Figure 3. Example images from scene database.

Table 1. Compared Results on Caltech dataset. * means no results in specific category.

Classes	Cars	Face	Airp.	Motor.	Leopard
Ours	98.218	96.393	97.74	91.924	100
R. Fergus	88.5	96.4	90.2	92.5	*
Opelt	83	93.5	88.9	92.2	*
D. Holub	*	91.00	93.8	95.1	93.00

gram of the proposed co-training method is show in Fig. 2.

4 Experimental results

Evaluation of the the proposed algorithm with two-view features is done using two different dataset: A scene dataset and object dataset, respectively. The first database is a collected database from Internet for auto-scene classification in digital camera system, which includes 8 categories (Beach, Snow, Flower, Cooking, Firework, Night landscape, Sunset and party). Each scene category include at less 300 images, and the total number of scene images is 8,235 (Different number for different categories). The sample images are shown in Fig. 3). For the scene database, The one-view feature is selected as color histogram, which is improved to be efficient for discriminating in scene-type images; the other is BOW feature, which is popular and effective for image understanding. In the experiments of scene classification, we randomly select 20 images from each scene class as the labeling training samples, and 1440 images as unlabeling samples, and then the remainder 6635 images as testing. In each co-training run, we train the SVM learner for each view using the current labeling samples, and select N candidate of training samples with high-confidence degree from each scene category. Then, if the probabilities of all selected candidate samples in one learner are larger than one selected threshold, we will update the these samples as the training examples of next run for the other learner. Otherwise, we compare the number, whose probabilities are larger than the threshold for the two learner, and update the candidate samples of the learners with large number of high probability as the training samples. We also fusion two learner for evaluating the classification rate with the two view features. Figure 4 give the experimental results with different view feature and fusion learner, which is evaluated with only labeled samples or with all training samples. It is obvious that the classification rate can be improved by using the unlabeled training samples with co-training method.

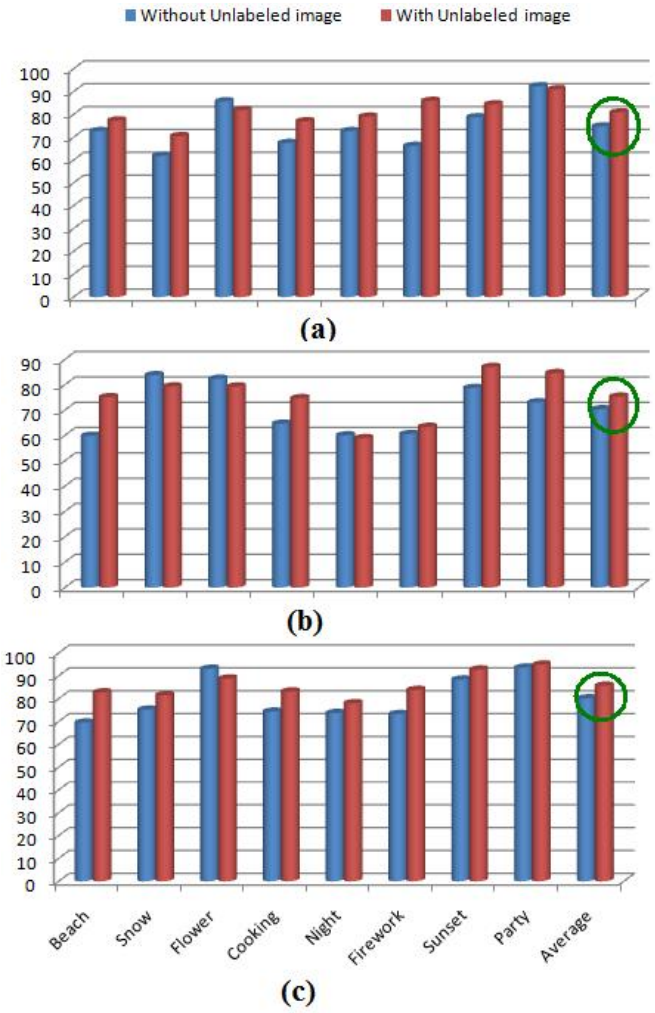

Figure 4. Compared results between with only label training samples and with all training samples for scene database; (a) Using color histogram; (b) using BOW feature; (c) Combining two learners.

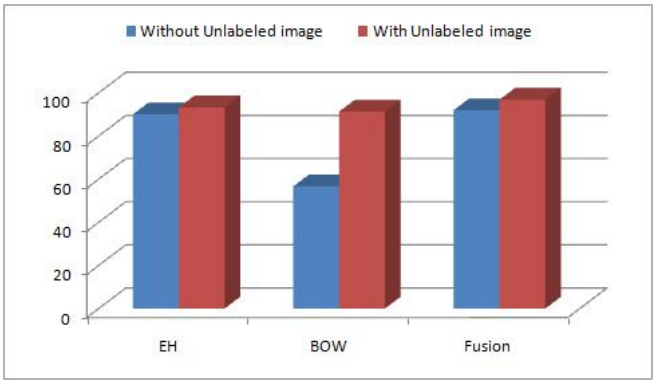

Figure 5. Compared results between with only label training samples and with all training samples for object database.

The other database is object datasets (Caltech Database[7]) by Rob Fergus and Pietro Perona to be evaluated with the experiment, which usually was tested by current remarkable works [5,6,7]. Experiments are performed with 5 classes object images(Car rear(1155), faces(450), airplanes(1074), motorbikes(826), leopards(200)). In the experiments, we randomly select 5 images as labeling training samples from each class, 750 images as unlabeling training samples, and the remainder are as testing images. For the scene database, The one-view feature is selected as edge histogram in regions, which is efficient for object-type images; the other is BOW feature. Fig. 5 gives the compared results with only labeled samples or with all training samples. Table 1 gives the compared results by our proposed algorithm with those of the state of art algorithm. It is obvious that the recognition performance by our proposed method can be greatly improved compared with those of the state of art algorithms for most object categories.

References

[1] Y.X. Chen, J.Z. Wang, "A region-based fuzzy feature matching approach to content-based image retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(9), 1252- 1267, 2002.

[2] Datta, Ritendra; Dhiraj Joshi, Jia Li, James Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys 40 (2): 1-60, Apr. 2008.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categoraization with bags of keypoints", In Proc. ECCVWorkshop on Statistical Learning in Computer Vision, pp. 1-16,

[4] I. Muslea, S. Minton, C.A. Knoblock, "Selective sampling with redundant views", In Proc. of the 17th National Conference on Artificial Intelligence, 2000, pp. 621–626.

[5] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning", In Proc. CVPR, 2003.

[6] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In Proc. ECCV, volume 2, pp. 7184, 2004.

[7] A.D. Holub, M. Welling, P. Perona "Hybrid Generative-Discriminative Object recognition". (International Journal of Computer Vision (IJCV), 77(1-3), 239-258, 2008.