Localized Content-Based Image Retrieval Through Evidence Region Identification

Wu-Jun Li & Dit-Yan Yeung
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong, China
{liwujun,dyyeung}@cse.ust.hk

Abstract

Over the past decade, multiple-instance learning (MIL) has been successfully utilized to model the localized content-based image retrieval (CBIR) problem, in which a bag corresponds to an image and an instance corresponds to a region in the image. However, existing feature representation schemes are not effective enough to describe the bags in MIL, which hinders the adaptation of sophisticated single-instance learning (SIL) methods for MIL problems. In this paper, we first propose an evidence region (or evidence instance) identification method to identify the evidence regions supporting the labels of the images (i.e., bags). Then, based on the identified evidence regions, a very effective feature representation scheme, which is also very computationally efficient and robust to labeling noise, is proposed to describe the bags. As a result, the MIL problem is converted into a standard SIL problem and a support vector machine (SVM) can be easily adapted for localized CBIR. Experimental results on two challenging data sets show that our method, called EC-SVM, can outperform the state-of-the-art methods in terms of accuracy, robustness and efficiency.

1. Introduction

1.1. Background

According to the low-level image features used in the retrieval process, existing *content-based image retrieval* (CBIR) methods can be categorized into two major classes, namely, global methods and localized methods (a.k.a. *localized CBIR* [11, 12]). Global methods exploit features characterizing the global view of an image, such as color histograms, to compute the similarity between images. These methods have been widely used by traditional CBIR systems. Although global features can be extracted easily, in many cases, only a small part or several small parts of the image are useful for characterizing the visual content of the image. If features from the whole image area are used to represent an image, the useful information may be overridden by noisy information from irrelevant regions. For example, in Figure 3, if the interest of the user is in the object "FabricSoftenerBox", the two images with label "FabricSoftenerBox" should have higher similarity than the first two images in the upper row. However, the first two images in the upper row are expected to give higher similarity than the two images in the leftmost column if global methods are used. On the contrary, localized CBIR [11, 12, 13], which describes the task where the user is only interested in a portion of the image with the rest being irrelevant, is more natural and is in line with human perception. For example, in Figure 3, a user may only be interested in the *apple* in the image with label "Apple".

A new learning paradigm called *multiple-instance learning* (MIL) [6][1] was proposed to model learning problems where the class labels are only associated with sets of examples rather than individual examples. In MIL, an individual example is called an *instance* and a *bag* contains a set of instances. Training labels are associated with bags rather than instances. A bag is labeled positive if at least one of its instances is positive; otherwise, the bag is negative. In this paper, we use the term *single-instance learning* (SIL) to refer to the traditional supervised learning paradigm in which each individual example has a class label.

In the existing localized CBIR work, the region of interest can be either at a fixed location or marked by the user. The first case does not conform to the general image retrieval task and the second case requires too much effort from the user, making it unappealing in practice. Hence, the *focus of this paper* is to design a general automatic localized CBIR system that does not necessarily require the user to mark the region of interest. Specifically, we require that multiple labeled images be provided for the system to auto-

[1] Due to the page limit constraint, in this paper, we can only cite the most related references from the computer vision community or those focused on vision applications. Many other references, especially those from the machine learning community, can be found in [7].

matically learn the interest of the user. This can be achieved through relevance feedback or by inputting a *query image set* [12] labeled as positive or negative by the user according to whether the images contain the target regions of interest. Under this setting, the underlying learning problem for localized CBIR is essentially an MIL problem where an image corresponds to a bag and each region in the image corresponds to an instance.

1.2. Motivation

Few of the existing MIL methods have designed effective feature representation schemes to describe the bags, making it difficult to adapt some sophisticated SIL methods for MIL problems. DD-SVM [4] is the first MIL method trying to propose a feature representation scheme for the bags in MIL to convert MIL into SIL. However, the features of DD-SVM are very sensitive to noise and incur very high computation cost. MILES [3] (Multiple Instance Learning via Embedded instance Selection) also converts MIL into a standard SIL problem via feature mapping, in which each feature is defined by an instance from the training bags, including both positive and negative bags. Although MILES is less sensitive to noise and more efficient than DD-SVM, the feature space for representing bags is of very high dimensionality because it contains too many irrelevant features. Hence, appropriate classifiers that can make use of the feature representation scheme in MILES are limited to those that can perform both feature selection and classification simultaneously, such as 1-norm SVM [3]. Therefore, *the motivation* of this work is to design an effective as well as efficient feature representation scheme for representing the bags in MIL.

1.3. Main Contributions

In this paper, we propose a feature representation scheme for the bags in MIL to convert MIL into SIL and adapt the sophisticated SIL technique, SVM, to solve MIL problems. The main contributions are summarized as follows:

- We propose an evidence region (or evidence instance) identification method to identify the evidence regions that support the labels of the images (i.e., bags).
- A very effective feature representation scheme, which is also very computationally efficient and robust to labeling noise, is proposed to describe the bags based on the identified evidence regions. As a result, the MIL problem is converted into a standard SIL problem and an SVM is successfully adapted for localized CBIR. The resulting method is called EC-SVM, which will be described in detail later.
- We compare our method extensively with many state-of-the-art methods on two challenging data sets to demonstrate the promising performance of our method with respect to multiple performance metrics, including accuracy, efficiency and robustness.

It should be emphasized that the *focus of this work* is on CBIR rather than image classification. Although the techniques for CBIR are also suitable for image classification, and vice versa, their application scenarios are somewhat different. While for image classification a large number of labeled images can be provided for training, for CBIR it is unreasonable (or impractical) to require the user to input a large number of query images.

2. A Feature Representation Scheme for MIL

2.1. Notations and Conventions

B_i^+ denotes a positive bag and B_i^- denotes a negative bag. When the label of a bag is irrelevant, we simply denote the bag as B_i. B_{ij}^+ denotes an instance in a positive bag B_i^+ and B_{ij}^- is an instance in a negative bag B_i^-. Let $\mathfrak{B} = \{B_1^+, B_2^+, \dots, B_{n^+}^+, B_1^-, B_2^-, \dots, B_{n^-}^-\}$ denote the set of all n^+ positive and n^- negative training bags. For each bag B_i, its bag label is $y_i \in \{+1, -1\}$. All the instances are represented as feature vectors of the same dimensionality. Furthermore, in CBIR, a bag refers to an image and an instance corresponds to a region in some image.

2.2. Evidence Instance Identification

According to the MIL problem formulation, a bag is labeled positive if at least one of its instances is positive; otherwise, the bag is labeled negative. Because whether or not there exist positive instances in a bag provides *evidence* for supporting the bag's label, we call the positive instances *evidence instances*. If a bag refers to an image, evidence instances are also referred to as *evidence regions*.

2.2.1 Evidence Instance Identification Algorithm

The *evidence confidence* $\mathrm{EC}(B_{gh})$, which is used to represent the confidence (or likelihood) for the instance B_{gh} to be an evidence instance, is defined as follows:

$$\mathrm{EC}(B_{gh}) = \prod_{i=1}^{n^+} \Pr(B_{gh} \mid B_i^+) \prod_{i=1}^{n^-} \Pr(B_{gh} \mid B_i^-), \quad (1)$$

where $\Pr(B_{gh} \mid B_i)$ is estimated based on the noisy-OR model [8]:

$$\Pr(B_{gh} \mid B_i^+) \quad \propto \quad \left\{1 - \prod_j \left[1 - \Pr(B_{gh} \mid B_{ij}^+)\right]\right\} (2)$$

$$\Pr(B_{gh} \mid B_i^-) \quad \propto \quad \prod_j \left[1 - \Pr(B_{gh} \mid B_{ij}^-)\right]. \quad (3)$$

Here, $\Pr(B_{gh} \,|\, B_{ij})$ is estimated as follows:

$$\Pr(B_{gh} \,|\, B_{ij}) \propto \exp\left\{ -\frac{\sum_k (B_{ijk} - B_{ghk})^2}{\sigma^2} \right\}, \qquad (4)$$

where σ is a scaling parameter, k ranges over all the features, and B_{ijk} and B_{ghk} refer to the kth features of the corresponding feature vectors.

The noisy-OR model conforms well to the MIL formulation. From (2), we can see that as long as one instance in B_i^+ is close to B_{gh}, $\Pr(B_{gh} \,|\, B_i^+)$ will be high. From (3), we can see that only if all the instances in B_i^- are far away from B_{gh}, $\Pr(B_{gh} \,|\, B_i^-)$ will be high. Hence, if every positive bag contains at least one instance close to B_{gh} and simultaneously all the instances in the negative bags are far away from B_{gh}, $\mathrm{EC}(B_{gh})$ will be high. Therefore, $\mathrm{EC}(\cdot)$ actually reflects the *confidence* for the instance to be an evidence instance. The larger the EC value of the instance, the more likely this instance will be an evidence instance.

The definition of EC "looks" similar to that of DD [8]. However, except that both EC and DD use the noisy-OR model to compute the corresponding probability, the rationales for EC and DD are in fact very different. This will be demonstrated in detail in the following subsection.

From the MIL definition, we know that evidence instances only exist in the positive bags and each positive bag contains at least one evidence instance. Hence, we just need to compute the EC values for all instances from the positive bags and then select those instances with the largest EC values from each positive bag to be our evidence instances.

Another issue about the above evidence instance identification method is how many evidence instances should be selected from each positive bag. This may be determined from prior knowledge. More specifically, for localized CBIR, this parameter can be completely observed from the given training images. For example, for the SIVAL image set [12, 13] used in our experiment, since from the training images we observe that the target object occupies about 15% of the image area for most images and each image (bag) contains 32 instances, it is very reasonable to set this parameter to 5 which is about 15.6% $(5/32)$ of the number of all instances in a bag.

Algorithm 1 summarizes the evidence instance identification procedure presented above.

Algorithm 1 Evidence Instance Identification for MIL

```
Input: All training bags $B_1^+, \ldots, B_{n^+}^+, B_1^-, \ldots, B_{n^-}^-$;
Parameter $m$ indicating how many evidence instances
should be identified from each positive bag.
Initialize: $E^* = \phi$
for $g = 1$ to $n^+$ do
    for $h = 1$ to $|B_g^+|$ do
        Compute $\mathrm{EC}(B_{gh}^+)$ according to (1)
    end for
    Select $m$ instances with the largest EC values from
    $B_g^+$, and add the selected instances to $E^*$
end for
Output: $E^*$, a set of identified evidence instances.
```

2.2.2 Comparison with DD

The DD method [8] tries to find the *target point* [2] by maximizing the following objective function:

$$\arg\max_c \prod_{i=1}^{n^+} \Pr(c \,|\, B_i^+) \prod_{i=1}^{n^-} \Pr(c \,|\, B_i^-), \qquad (5)$$

where $c \in C$ and C is the space of all possible instances, including both the observed training instances in $\mathfrak{B}$ and the (possibly infinite number of) unobserved ones.

$\Pr(c \,|\, B_i)$ is also estimated based on the noisy-OR model [8]. However, unlike our EC definition, $\Pr(c \,|\, B_{ij})$ in DD is estimated as follows:

$$\Pr(c \,|\, B_{ij}) \propto \exp\left\{ -\sum_k \left(s_k (B_{ijk} - c_{\cdot k})^2 \right) \right\}, \qquad (6)$$

where c corresponds to a feature vector, which might not be an observed instance, in the input instance space, k ranges over all the features, s_k is a scaling coefficient for the kth feature, and B_{ijk} and $c_{\cdot k}$ refer to the kth features of the corresponding feature vectors.

The main difference between EC and DD can be easily seen from the difference between (1) and (5), where the EC value, which is defined only for the *observed* instances in $\mathfrak{B}$, can be directly computed from the training data, while DD tries to *maximize* an objective function, i.e., to search for the target point, over C which is a continuous space with infinitely many members. The flow charts of EC computation and DD are illustrated in Figure 1 and Figure 2, respectively. In Figure 1, the *direct computation* step is based on (1) without the need for any optimization procedure. In Figure 2, however, an *optimization procedure*, such as gradient ascent in [8], should be firstly applied to find the target point c_t by maximizing the objective function in (5). Then, based on c_t, a value, such as the distance between B_{gh} and c_t in [8], is computed by the *further computation* step for further processing.

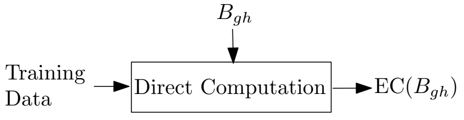

Figure 1. Flow chart for the evidence confidence (EC) computation.

[2] This target point is not necessarily an observed instance in the training set $\mathfrak{B}$. We must search for it in the whole instance space which may be a continuous space containing infinitely many instances.

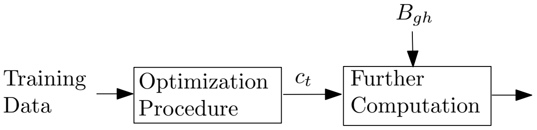

Figure 2. Flow chart for the diverse density (DD) method.

From (5), it is not difficult to realize that the *optimization procedure* in DD is very sensitive to labeling noise. For example, if we mislabel just a single positive bag $\hat{B}$ as a negative bag, then $\Pr(c_t \mid \hat{B})$ computed based on (3) for the true target point c_t will decrease exponentially. As a result, the objective function value on c_t is likely to be very small. Hence, in this case, the computed target point will be relatively far away from the true target point. This problem has been validated by the experiments in [3, 4]. Moreover, the DD landscape typically contains local maxima. Searching for the true target point by applying gradient-ascent or EM does not guarantee global optimality. With no prior knowledge for a good initialization point, multiple restarts are generally needed and hence high computation cost is incurred.

Another difference between EC and DD comes from the difference between (4) and (6). In (4), all the features (indexed by k) have the same scaling parameter σ. In (6), each feature (k) has its specific scaling parameter s_k. The adoption of (4) for EC computation is motivated by MILES [3]. In MILES, they use the same scaling parameter for all the features, but the performance of MILES is still better than DD-SVM [4] which adopts a scaling vector to weight the features. The computation cost for the EC value based on (4) will be dramatically decreased because we do not have to search through a huge space of possible scaling coefficient values. Moreover, the meaning of $\Pr(B_{gh} \mid B_{ij})$ in (4) is much more obvious, which is just a kernel density estimate with B_{ij}.

Our evidence instance identification method based on EC computation *totally avoids* the two disadvantages of DD-based methods, which are high computation cost and high sensitivity to labeling noise. The advantages of our method are summarized as follows:

- The EC value of each *observed instance* is directly computed from the training set. The parameter m in Algorithm 1 can be obtained from prior knowledge or observed directly from the training data for image retrieval. In our experiments, we just set σ to 1 if the data are normalized, and the performance is still very promising. Hence, our method essentially has no parameters to tune, making it several orders of magnitude faster than DD-based methods.

- Because our evidence instance identification method constrains the search scope for identifying evidence instances to be within a bag, it is also very robust towards labeling noise. To illustrate this, let us assume that x^+ is a true positive instance from a positive training bag and x^- is a negative instance (false positive instance) from the same bag. Without labeling noise, most (or even all) terms, $\Pr(x^+|B_i^+)$ and $\Pr(x^+|B_i^-)$, should be expected to be larger than their counterparts, $\Pr(x^-|B_i^+)$ and $\Pr(x^-|B_i^-)$, in (1). Hence, $\text{EC}(x^+)$ should be much larger than $\text{EC}(x^-)$. Even if a portion of the training bags are mislabeled to make *some* terms, $\Pr(x^-|B_i^+)$ and $\Pr(x^-|B_i^-)$, larger than their counterparts, $\Pr(x^+|B_i^+)$ and $\Pr(x^+|B_i^-)$, $\text{EC}(x^+)$ will still be larger than $\text{EC}(x^-)$ as long as the number of these terms is not too large. Even if $\text{EC}(x^+)$ will decrease in this case, it will not affect the evidence instance identification result because only the *relative EC values*, rather than the absolute EC values, for the instances in a specific bag will affect the result in Algorithm 1.

2.3. Feature Representation Scheme

Based on the identified evidence instances, we propose a feature mapping to map every bag B_i to a point $\psi(B_i)$ in the evidence instance based feature space:

$$\psi(B_i) = \Big(d(e_1^*, B_i), d(e_2^*, B_i), \ldots, d(e_{|E^*|}^*, B_i)\Big)^T, \quad (7)$$

where $e_k^* \in E^*$, E^* is the set of identified evidence instances in Algorithm 1, and $d(e, B_i)$ is defined as follows:

$$d(e, B_i) = \min_{B_{ij} \in B_i} (\|e - B_{ij}\|^2), \quad (8)$$

which means the distance between an instance and a bag is simply equal to the distance between the instance and the nearest instance in the bag.

This feature mapping is very meaningful because generally the distance between two evidence instances is expected to be smaller than the distance between one evidence instance and a non-evidence instance from the background. Because positive bags contain evidence instances, the distance from one evidence instance to a positive bag is expected to be smaller than the distance from this evidence instance to a negative bag. Hence, the features in (7) are expected to have strong discrimination ability. Furthermore, for a specific bag, different instances in it will be selected as the nearest instances to compute the distance in (8) for different e_k^*. Hence, the feature vector in (7) actually implicitly contains the inter-dependency between the instances in a bag, the effectiveness of which has been validated by [9].

3. Single Instance Formulation for MIL

After the feature mapping defined in (7), the MIL problem is converted into a standard SIL problem and hence any conventional classification method can easily be adapted for the MIL problem.

In this paper, we adapt SVM for MIL because it can deliver promising generalization performance via margin maximization. The resulting method is called EC-SVM. Since SVM has become a mature technique which has been widely used in many applications, we do not introduce it in detail here. We refer the readers to the related literature, such as LIBSVM [2] and its documentation.

4. Relation to Existing Work

From the formulation point of view, EC "looks" similar to DD. However, EC's modification to DD makes EC-SVM ingeniously integrate the advantages of both MILES and DD-SVM, and simultaneously overcome their shortcomings. From MILES, we can see that using the instances from the training bags to construct the feature representation is sufficient for good performance. Hence, the optimization procedure in DD for finding local optima, which is both time-consuming and noise sensitive, is unnecessary. From DD-SVM, we can see that the most discriminative features might be those constructed based on evidence instances. Hence, the features constructed based on negative instances, which are adopted by MILES, might be useless, or even harmful (cf. Figure 7 and the discussion). EC-SVM constructs only those discriminative features corresponding to evidence instances without any time-consuming optimization procedure. Hence, EC's modification to DD makes EC-SVM much more effective than MILES and DD-SVM.

5. Performance Evaluation

We evaluate EC-SVM based on two publicly available image data sets: the SIVAL (Spatially Independent, Variably Area, and Lighting) image set [12, 13] and the COREL image set [3]. As for SVM training, the Gaussian kernel $\kappa(x,y) = \exp^{-r\|x-y\|^2}$ is used for our method in all the experiments. We use LIBSVM [2] to train all the SVM classifiers.

Note that the motivation of our paper is to design an effective feature representation scheme to describe the bags. Hence, among all the proposed MIL methods, MILES and DD-SVM are the most related methods. Because previous work [3] has shown that MILES outperforms DD-SVM, MILES is adopted as the baseline for EC-SVM. We only compare EC-SVM with DD-SVM in terms of computational cost and noise sensitivity to verify the claims in Section 2.2.2.

5.1. Accuracy Evaluation

5.1.1 Evaluation on SIVAL Data Set

The SIVAL data set contains 1,500 images of 25 categories, with 60 images for each category. Category 1 to category 25 are: "AjaxOrange", "Apple", "Banana", "BlueScrunge", "CandleWithHolder", "CardboardBox", "CheckeredScarf", "CokeCan", "DataMiningBook", "DirtyRunningShoe", "DirtyWorkGloves", "FabricSoftenerBox", "FeltFlowerRug", "GlazedWoodPot", "GoldMedal", "GreenTeaBox", "JuliesPot", "LargeSpoon", "RapBook", "SmileyFaceDoll", "SpriteCan", "StripedNoteBook", "TranslucentBowl", "WD40Can", and "WoodRollingPin". Figure 3 shows some sample images from the SIVAL image set. We use the same preprocessing method as that in [10, 12] to generate the bags. Hence, each image is represented as a bag of 32 30-dimensional instances.

FabricSoftenerBox CheckeredScarf Apple

FabricSoftenerBox CheckeredScarf SpriteCan

Figure 3. Sample images from the SIVAL image set.

We compare the performance of EC-SVM with several related methods on this data set. Note that for fair comparison, we just list the results of the methods that adopt the same bag generation method. For example, for MI-Winnow [5], there are two bag generation methods, called "Neighbors" and "No Neighbors" respectively. The "Neighbors" method is the same as that introduced in this paper. We just list the results of MI-Winnow with the "Neighbors" bag generation method.

We adopt the same experimental settings as those used by related methods. For each category, we use the "one-versus-the-rest" strategy to evaluate the performance. We randomly select eight positive and eight negative images to form the training set and the remaining 1,484 images to form the test set. Unless stated otherwise, the results are reported based on 30 rounds of independent test. Because the target object occupies about 15% of the image area for most images, we simply set the parameter m in Algorithm 1 to 5 which is about 15.6% $(5/32)$ of the number of all instances in a bag. The parameter C and Gaussian kernel parameter r for SVM in LIBSVM [2] are simply set to 1 and 2^{-4} respectively. Better performance might be expected if a more sophisticated method, such as cross-validation on the training data, is used to set these parameters. For the parameters in MILES [3], we find that $\lambda = 0.2$ and $\sigma^2 = 1$ give the best *test performance* for the SIVAL data set. Hence, we fix $\lambda = 0.2$ and $\sigma^2 = 1$ for MILES in all the following experiments. The average AUC (area under the ROC curve) values with 95% confidence interval for the 25 categories are reported in Table 1, in which ACCIO! is introduced in [12]. We can see that EC-SVM achieves the best performance for most categories.

Table 1. Average AUC values (in percent) with 95% confidence interval over 30 rounds of test on the SIVAL image set. The best performance is shown in bold.

Category ID	EC-SVM	MILES	MI-Winnow	ACCIO!
1	**93.8** $\pm$ **2.1**	90.2 ± 2.3	83.0 ± 3.6	77.0 ± 3.4
2	**68.0** $\pm$ **2.6**	64.5 ± 2.5	58.5 ± 5.9	63.4 ± 3.4
3	**69.1** $\pm$ **2.9**	68.1 ± 3.1	59.8 ± 3.1	65.9 ± 3.3
4	**74.1** $\pm$ **2.4**	72.6 ± 2.5	58.6 ± 5.1	69.5 ± 3.4
5	**88.1** $\pm$ **1.1**	84.0 ± 2.3	86.1 ± 1.5	68.8 ± 2.3
6	**85.6** $\pm$ **1.6**	81.2 ± 2.7	72.5 ± 3.8	67.9 ± 2.2
7	**96.9** $\pm$ **0.5**	93.7 ± 1.2	93.2 ± 1.2	90.8 ± 1.6
8	**94.6** $\pm$ **0.8**	92.4 ± 0.8	91.9 ± 2.4	81.5 ± 3.5
9	**75.0** $\pm$ **2.4**	71.1 ± 3.2	74.5 ± 4.5	74.7 ± 3.4
10	**90.3** $\pm$ **1.3**	85.3 ± 1.7	84.4 ± 1.7	83.7 ± 1.9
11	**83.0** $\pm$ **1.3**	77.1 ± 3.1	72.0 ± 3.1	65.3 ± 1.5
12	**97.9** $\pm$ **0.5**	97.1 ± 0.7	95.6 ± 1.1	86.6 ± 3.0
13	**94.2** $\pm$ **0.8**	93.9 ± 0.7	88.7 ± 1.5	86.9 ± 1.7
14	68.0 ± 2.8	68.2 ± 3.1	58.5 ± 3.0	**72.7** $\pm$ **2.3**
15	**87.5** $\pm$ **1.4**	80.7 ± 2.9	74.1 ± 4.9	77.7 ± 2.6
16	86.9 ± 2.2	**91.2** $\pm$ **1.7**	86.4 ± 3.0	87.3 ± 3.0
17	67.3 ± 3.3	78.7 ± 2.9	72.1 ± 5.8	**79.2** $\pm$ **2.6**
18	**61.3** $\pm$ **1.8**	58.2 ± 1.6	52.9 ± 2.5	57.6 ± 2.3
19	**68.6** $\pm$ **2.3**	61.7 ± 2.4	58.3 ± 3.1	62.8 ± 1.7
20	**84.6** $\pm$ **1.9**	77.5 ± 2.6	72.4 ± 3.8	77.4 ± 3.3
21	85.4 ± 1.2	80.4 ± 2.0	**85.6** $\pm$ **1.8**	71.9 ± 2.5
22	**75.6** $\pm$ **2.3**	68.7 ± 2.4	72.4 ± 3.8	70.2 ± 3.2
23	74.2 ± 3.2	73.2 ± 3.1	70.4 ± 5.3	**77.5** $\pm$ **2.3**
24	**94.3** $\pm$ **0.6**	88.1 ± 2.2	90.7 ± 1.4	82.0 ± 2.4
25	**66.9** $\pm$ **1.7**	62.1 ± 2.5	57.0 ± 2.9	66.7 ± 1.7
Average	**81.3**	78.4	74.8	74.6

We further test EC-SVM by varying the size of the training set. The average AUC values for all 25 categories over 30 rounds of test, together with the results reported by other methods, are listed in Table 2, in which the first row shows the number of training images for each class. For example, the number "1" refers to the case in which one positive image and one negative image are selected for training and all other images for testing. We can see that EC-SVM achieves the best performance for all cases.

Table 2. Average AUC values (in percent) for all 25 categories over 30 rounds of test on the SIVAL image set. N/A denotes the case in which the corresponding method did not report results for that setting.

	1	2	4	8	12
MISSL [10]	N/A	N/A	N/A	74.8	N/A
MI-Winnow	N/A	N/A	66.8	74.8	79.4
MILES	58.7	64.5	71.7	78.4	82.0
EC-SVM	**66.0**	**70.1**	**76.0**	**81.3**	**84.2**

5.1.2 Evaluation on COREL Data Set

As in MILES [3], we choose 2,000 images from 20 (category 0 to category 19) COREL Photo CDs. Each CD contains 100 images representing a different category. The images are in JPEG format with size 384×256 or 256×384. We use the same image segmentation and feature representation methods in MILES to construct the corresponding bags and instances. After segmentation, each region in an image is characterized by a nine-dimensional feature vector representing the color, texture and shape information from the region. Figure 4 shows one sample image from each of the 20 categories. The categories are ordered in a row-wise manner from the upper-leftmost image (category 0) to the lower-rightmost image (category 19).

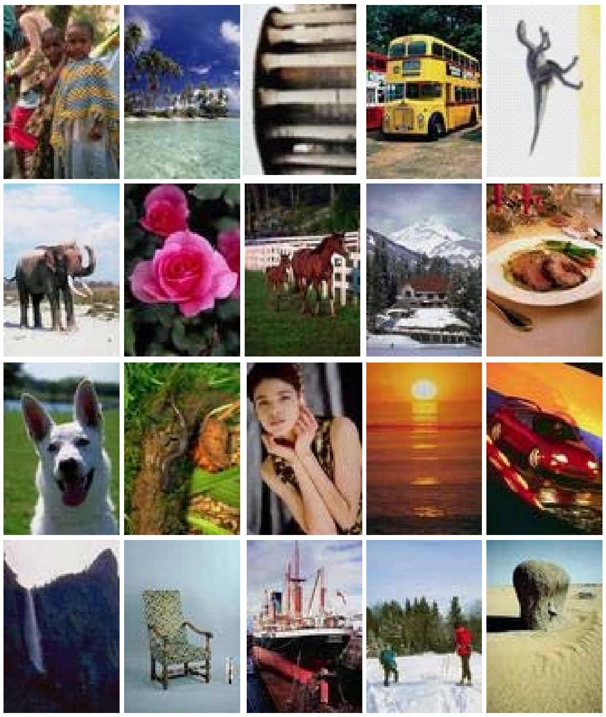

Figure 4. Sample images from the 20 categories of the COREL image set.

Because MILES has achieved better results than many other methods [3], including both global methods and local methods, we choose MILES as the *baseline* for comparison. As in [3], we choose λ from 0.1 to 0.6 with step size 0.05 and σ^2 from 5 to 15 with step size 1. We find that $\lambda = 0.2$ and $\sigma^2 = 11$ give the best *test performance* for MILES on the COREL data set. Hence, we fix $\lambda = 0.2$ and $\sigma^2 = 11$ for MILES in all the following experiments.

For each category, we use the "one-versus-the-rest" strategy to evaluate the performance. In each round, $n \in \{1, 2, 4\}$ randomly selected positive images and n randomly selected negative images are chosen to form the training set and the remaining $2,000 - 2n$ images to form the test set. The results are reported based on 50 rounds of independent test. Although the target objects in different categories, or the target objects from the same category but in different images, may be partitioned into different number of regions, we simply set the parameter m in Algorithm 1 to 3. The parameter C and Gaussian kernel parameter r for SVM in LIBSVM [2] are set to 1 and 2^{-3} respectively. Table 3 lists the results of the average AUC values (in percent) for all 20 categories with 95% confidence interval over 50 rounds of test. Once again, EC-SVM achieves better results than MILES.

Table 3. Average AUC values (in percent) for all 20 categories with 95% confidence interval over 50 rounds of test on the COREL image set.

n	1	2	4
MILES	64.4 ± 1.1	72.2 ± 0.8	79.6 ± 0.6
EC-SVM	**76.4± 0.6**	**80.0±0.4**	**83.2±0.3**

Figure 5 shows the average AUC values with 95% confidence interval for each category when $n = 1$.

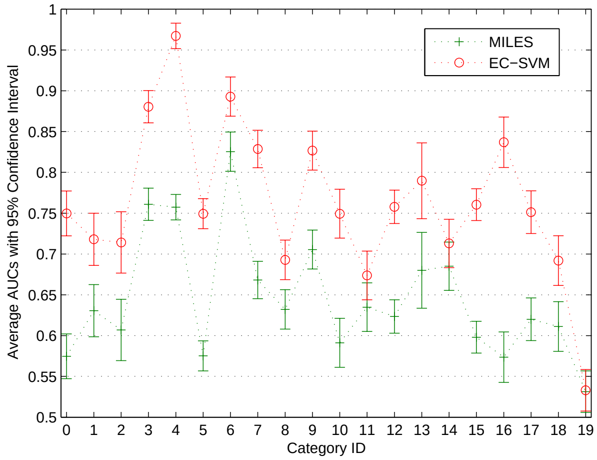

Figure 5. Comparison on the COREL data set with one positive and one negative examples labeled.

5.2. Sensitivity to Labeling Noise

We use the same setting as that in MILES [3] to evaluate the noise sensitivity on the COREL data set. We add $d\%$ of noise by changing the labels of $d\%$ of positive bags and $d\%$ of negative bags. We compare EC-SVM with DD-SVM and MILES under different noise levels based on 200 images from Category 2 ("Historical buildings") and Category 7 ("Horses"). The training and test sets are of the same size. The average classification accuracy over five randomly generated test sets is shown in Figure 6. We can see that MILES and EC-SVM are much more robust than DD-SVM, and the robustness of EC-SVM is comparable with MILES.

We further test the noise sensitivity of EC-SVM on the SIVAL data set. We compare EC-SVM with MILES under different noise levels ($n/30, n = 1, \ldots, 9$), by negating the labels of n positive and n negative training images, based on 120 images from Category 7 ("CheckeredScarf") and Category 12 ("FabricSoftenerBox"). The training and test sets are of the same size. The average classification accuracy with 95% confidence interval over 30 randomly generated test sets is shown in Figure 7. We can see that EC-SVM is much more robust than MILES on the SIVAL data set.

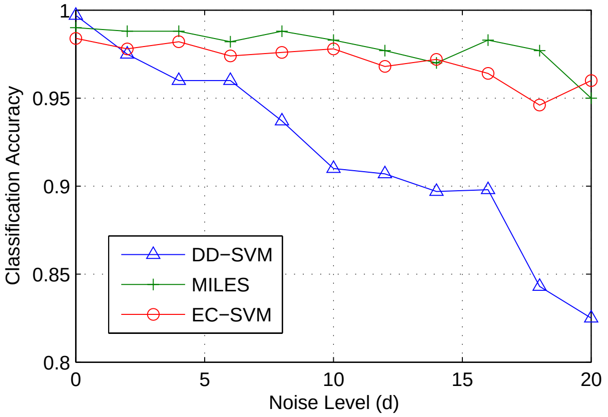

Figure 6. Comparison of sensitivity to labeling noise on the COREL data set.

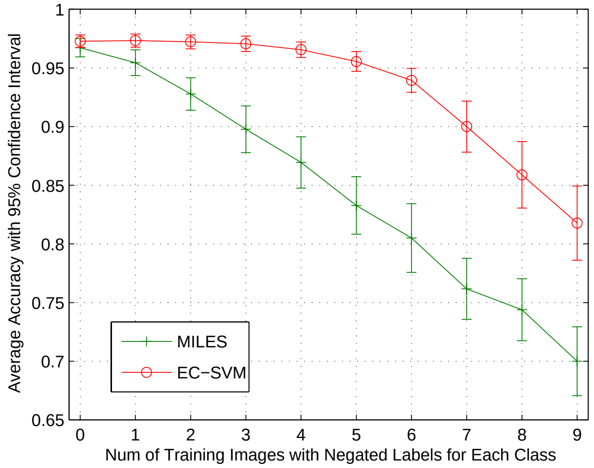

Figure 7. Comparison of sensitivity to labeling noise on the SIVAL data set.

The SIVAL data set differs from the COREL data set in many aspects. In COREL, the target object occupies a large portion of the whole image, while in SIVAL the main part of an image is the background. Furthermore, the background of some category in COREL is always specific to that category of images. For example, in general, the background in the images of "Historical buildings" (Category 2) is very different from the background in the images of "Horses" (Category 7), which can be easily seen from Figure 4. But for SIVAL, the background for one category can appear for another category. MILES uses all the instances, from both positive training bags and negative training bags, as the basis for feature construction [3]. This will make the effect of instances from the background dominate the effect of the evidence instances on SIVAL. Because the background can appear in either positive or negative bags, the features based on instances from the background actually have very low discrimination ability. Hence, the useful features in MILES are very limited. As a result, MILES will be more easily affected by noise on the SIVAL data set. This might be the cause for the phenomenon that MILES is much more sensitive to noise on the SIVAL data set.

5.3. Computation Cost

Table 4 lists the training time (on a 2GHz PC with 1G memory) required by DD-SVM, MILES, and EC-SVM. "SIVAL" refers to the time for training 25 classifiers for all the 25 categories when four positive and four negative images are used as the training set on the SIVAL data set. "COREL" refers to the time for training 20 classifiers for all the 20 categories when four positive and four negative images are used as the training set on the COREL data set. To test the scalability of EC-SVM, we also evaluate the training time on the COREL data set based on a training set of 500 images, denoted as "COREL2", which has been used for efficiency comparison in MILES [3]. We can see that EC-SVM is much more efficient.

Table 4. Computation time comparison (in minutes).

	SIVAL	COREL	COREL2
DD-SVM	N/A	N/A	40
MILES	0.34	0.064	0.85
EC-SVM	**0.23**	**0.005**	**0.2**

6. Conclusion and Future Work

Considering the high computation cost and high noise sensitivity of DD-SVM, and the very high dimensionality of the feature vectors used by MILES, the feature representation scheme proposed in this paper is a much more practical one to effectively describe the bags in MIL.

Although very promising performance has been achieved by our method even though we simply use prior knowledge to determine how many evidence instances should be identified from each positive bag, a better choice is to learn this parameter from data. Different positive bags might have different numbers of evidence instances. Hence, how to adaptively identify the appropriate number of evidence instances for each positive bag will be pursued in our future work.

Furthermore, in CBIR, it is easy to get a large number of unlabeled images from the image repository. Hence, semi-supervised learning methods, which can incorporate unlabeled data into the training process, are very meaningful for CBIR. This will also be pursued in our future work. For example, we can apply manifold regularization [1] for semi-supervised localized CBIR.

Acknowledgements

This research has been supported by General Research Fund 621407 from the Research Grants Council of the Hong Kong Special Administrative Region, China. We thank Dr. Yixin Chen for sharing the code and data for MILES.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006. 8

[2] C.-C. Chang and C.-J. Lin. *LIBSVM: a Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm. 5, 7

[3] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, 2006. 2, 4, 5, 6, 7, 8

[4] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004. 2, 4

[5] S. R. Cholleti, S. A. Goldman, and R. Rahmani. Mi-Winnow: A new multiple-instance learning algorithm. In *18th IEEE International Conference on Tools with Artificial Intelligence*, pages 336–346, 2006. 5

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997. 1

[7] W.-J. Li and D.-Y. Yeung. MILD: Multiple-instance learning via disambiguation. *IEEE Transactions on Knowledge and Data Engineering*, In Press. 1

[8] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 1997. 2, 3

[9] G.-J. Qi, X.-S. Hua, Y. Rui, T. Mei, J. Tang, and H.-J. Zhang. Concurrent multiple instance learning for image categorization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. 4

[10] R. Rahmani and S. A. Goldman. MISSL: multiple-instance semi-supervised learning. In *Proceedings of the Twenty-Third International Conference Machine Learning*, pages 705–712, 2006. 5, 6

[11] R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts. Localized content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1902–1912, 2008. 1

[12] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts. Localized content based image retrieval. In *Multimedia Information Retrieval*, pages 227–236, 2005. 1, 2, 3, 5, 6

[13] H. Zhang, R. Rahmani, S. R. Cholleti, and S. A. Goldman. Local image representations using pruned salient points with applications to CBIR. In *ACM Multimedia*, pages 287–296, 2006. 1, 3, 5