Improving Image Categorization by Using Multiple Instance Learning with Spatial Relation

Thanh Duc Ngo[1,2], Duy-Dinh Le[2,1], and Shin'ichi Satoh[2,1]

[1] The Graduate University for Advanced Studies (Sokendai), Japan
[2] National Institute of Informatics, Tokyo, Japan
{ndthanh,ledduy,satoh}@nii.ac.jp

Abstract. Image categorization is a challenging problem when a label is provided for the entire training image only instead of the object region. To eliminate labeling ambiguity, image categorization and object localization should be performed simultaneously. Discriminative Multiple Instance Learning (MIL) can be used for this task by regarding each image as a bag and sub-windows in the image as instances. Learning a discriminative MI classifier requires an iterative solution. In each round, positive sub-windows for the next round should be selected. With standard approaches, selecting only one positive sub-window per positive bag may limit the search space for global optimum; meanwhile, selecting all temporal positive sub-windows may add noise into learning. We select a subset of sub-windows per positive bag to avoid those limitations. Spatial relations between sub-windows are used as clues for selection. Experimental results demonstrate that our approach outperforms previous discriminative MIL approaches and standard categorization approaches.

Keywords: Image Categorization, Multiple Instance Learning, Spatial Relation.

1 Introduction

We investigated image categorization using Multiple Instance Learning (MIL). Image categorization is a challenging problem especially when a label is provided for a training image only instead of the object region. Low categorization accuracy may result because the object region and background region within one training image share the same object label. To eliminate labeling ambiguity, image categorization and object localization should be simultaneously performed. In order to do that, one can use MIL, which is a generalization of standard supervised learning. Unlike standard supervised learning in which the training instances are definitely labeled, in the MIL setting, labels are only available for groups of instances called bags. A bag is positive if it contains at least one positive instance. Meanwhile, all instances in negative bags must be negative. Given training bags and instances that satisfy MIL labeling constraints, MIL approaches can learn to classify unlabeled bags as well as unlabeled instances

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 108–117, 2011.

in the bags. Thus, if we regard each image as a bag and sub-windows in images as instances, we can perform image categorization and object localization simultaneously using MIL.

Several MIL approaches have been proposed [1,2,3,4,5,6,7]. Empirical studies [2,4,7] demonstrate that generative MIL approaches perform worse than discriminative MIL approaches on benchmark datasets, because of their strict assumption on compact clusters of positive instances in the feature space. Thus, it is more appealing to tackle image categorization by using discriminative MIL approaches. In a brief overview, discriminative MIL approaches can be found in [5,4,6,7]. Andrews et al. [5] introduce a framework in which MIL is considered in different maximum margin formulations. A similar formulation of [5] can be found in [9]. DD-SVM presented in [4] trains an SVM for bags in a new feature space constructed from a mapping model defined by the local extremums of the Diverse Density function on instances of positive bags. In contrast, MILES [6] uses all instances in all training bags to construct the mapping model without applying any instance selection method explicitly. IS-MIL [7] then propose an instance selection method to tackle large-scale MIL problems. Because [4,6,7] heavily rely on bag-instance mapping process which is out of scope, we address our work to the framework proposed in [5].

In this paper, we extend the framework in [5] using spatial relations between sub-windows. Although spatial relation information have shown their important role in computer vision tasks [10,11,13,14], there is a few of MIL works utilizing such information. Zha et al. [12] introduced a MIL approach which captures the spatial configuration of the region labels. However, their work target to multi-label MIL problem and spatial relations between segmented regions. Instead of that, we investigate single-label MIL problem and overlapping relations between sub-windows. In the framework [5], learning a discriminative MI classifier is formulated as a non-convex problem and requires an iterative solution. In each round, positive training sub-windows (i.e. instances) for the next round should be selected with certain criteria. With original criteria, selecting only one positive sub-window per positive bag may limit the search space for the global optimum; meanwhile, selecting all temporal positive sub-windows may add noise into learning. We propose to select a subset of sub-windows per positive bag to avoid those limitations. Spatial relations between sub-windows are used as clues for selection. We directly enforce sub-windows spatial relations into learning by selecting sub-windows of the subset based on their overlapping degree with the most discriminative sub-window. Experimental results demonstrate the effectiveness of our approach.

2 Support Vector Machine for Multiple Instance Learning

In statistical pattern recognition, given a set of labeled training instances coupled with manual labels $(x_i, y_i) \in \mathcal{R}^d \times \mathcal{Y}$, the problem is how to obtain a classification function going from instances to labels $f : \mathcal{R}^d \rightarrow \mathcal{Y}$. In the binary case,

$\mathcal{Y} = \{-1, 1\}$ indicates positive or negative labels associated with instances. MIL generalizes this problem by relaxing the assumption on instance labeling. Labels are given for bags, which are groups of instances. A bag is assigned a positive label if and only if at least one instance of the bag is positive. Meanwhile, a bag is negative if all instances of the bag are negative. Formally, given a set of input instances $x_1, \ldots, x_n$ grouped into non-overlapping bags $B_1, \ldots, B_m$, with $B_I = \{x_i : i \in I\}$ and index sets $I \subseteq \{1, \ldots, n\}$. Each bag B_I is then given a label Y_I. Labels of bags are constrained to express the relation between bag and instances in the bag as follows: if $Y_I = 1$ then at least one instance $x_i \in B_I$ has label $y_i = 1$, otherwise, if $Y_I = -1$ then all instances $x_i \in B_I$ are negative: $y_i = -1$. A set of linear constraints can be used to formulate the relation between bag labels Y_I and instance labels y_i:

$$\sum_{i \in I} \frac{y_i + 1}{2} \geq 1, \forall I : Y_I = 1 \quad \text{and} \quad y_i = -1, \forall I : Y_I = -1, \tag{1}$$

or compactly represented as: $Y_I = \max_{i \in I} y_i$.

Learning the discriminative classifiers entails finding a function $f : \mathcal{X} \to \mathcal{R}$ for a multiple-instance dataset with the constraint $Y_I = \operatorname{sgn} \max_{i \in I} f(x_i)$.

3 The Former Approaches of SVM-Based Multiple Instance Learning

Andrews et al. [5] proposed two learning approaches based on SVM with different margin notions. The first approach, called mi-SVM, aims at maximizing the instance margin. Meanwhile, the second approach, called MI-SVM, tries to maximize the bag margin. Both mi-SVM and MI-SVM can be formed as mixed integer quadratic programs and need heuristic algorithms to be solved. The algorithms have an outer loop and an inner loop. The outer loop sets the values for the integer variables. Meanwhile, the inner loop trains a standard SVM. The outer loop stops if none of the integer variables changes in consecutive rounds.

The mixed integer formulation of mi-SVM based on the generalized soft-margin SVM can be presented as:

$$\begin{aligned} \min_{\{y_i\}} \min_{\{w,b,\xi\}} \quad & \frac{1}{2}\|w\|^2 + C \sum_i \xi_i \\ \text{subject to} \quad & \forall i : \; y_i \left(\langle w, x_i \rangle + b \right) \geq 1 - \xi_i \, , \, \xi_i \geq 0 \, , \\ & y_i \in \{-1, 1\} \text{ , and (1) hold.} \end{aligned} \tag{2}$$

In (2), labels y_i of instances x_i not belonging to any negative bag are treated as unknown integer variables. The target here is to find a linear discriminative *MI-separating* that satisfies the constraint wherein at least one positive instance from each positive bag lies in the positive half-space, while all instances belonging to all negative bags are in the negative half-space.

In MI-SVM, Andrews et al. introduce an alternative approach to the MIL problem. The notion of a margin is extended from individual instances to bags.

The margin of a positive bag is defined as the margin of *"the most positive"* instance of the bag. Meanwhile, the margin of a negative bag is defined by the margin of *"the least negative"* instance of the bag. Let $x_{mm(I)}$ be the instance of bag B_I and has maximum margin to the hyper-plane. Then, MI-SVM can be formulated as follows:

$$\begin{aligned} \min_{\{y_i\}} \min_{\{w,b,\xi\}} \quad & \frac{1}{2}\|w\|^2 + C\sum_I \xi_I \\ \text{subject to} \quad & \forall I:\ Y_I = -1 \wedge -\langle w, x_i\rangle - b \geq 1 - \xi_I\,,\ \forall i \in I\,, \\ & \text{or} \qquad Y_I = 1 \wedge \langle w, x_{mm(I)}\rangle + b \geq 1 - \xi_I, \text{ and } \xi_I \geq 0 \end{aligned} \tag{3}$$

4 Support Vector Machine with Spatial Relation for Multiple Instance Learning

MI-SVM and mi-SVM can be applied to image categorization by regarding each image as a bag and sub-windows in images as instances. However, their formulations and heuristic solutions do not involve spatial relations of sub-windows despite such information being extremely meaningful. Surrounding sub-windows always contain highly related information with respect to visual perception. If a sub-window in image is classified as a positive instance, it is supposed to be associated with the object label given to the class. In that sense, its neighboring sub-windows should be positive also. For example, if a sub-window tightly covers an object, its slightly surrounding sub-windows also contain that object.

Moreover, in terms of learning, the original approaches require a heuristic iterative solution to obtain the final discriminative classifier. In each learning round, candidate positive instances must be selected for the next round. Thus, positive instance selection criterion is the key step in the learning process. With mi-SVM, selecting all positive instances in the current round may add noisy instances to learning. Meanwhile, selecting only the most positive instance which has largest margin in the current round, as in MI-SVM, may limit the search space for the global optimum. To avoid such limitations, we propose to select a subset of instances as candidate positive instances for the next learning round. Spatial relations between instances (i.e. sub-windows) can be used as clues for selection. Therefore, we extend the framework proposed by Andrews et al. to take the spatial relation between sub-windows into account. Positive candidate selection criteria of the approaches are illustrated in Figure 1 .

In our extension, the notion of a bag margin is used as in the MI-SVM formulation. This means the margin of a positive bag is defined as the margin of *"the most positive"* instance of the bag. However, we directly enforce the spatial relations between *"the most positive"* instance with its spatially surrounding instances by adding constraints to the optimization formulation. Here, let $x_{mm(I)}$ be the instance of bag B_I has maximum margin with respect to the hyper-plane, and $\mathcal{SR}(x_{mm(I)}, T)$ denotes the set of $x_{mm(I)}$ and instances that surround $x_{mm(I)}$ with respect to the overlap parameter T. An instance belongs to $\mathcal{SR}(x_{mm(I)}, T)$ if its overlap degree with $x_{mm(I)}$ is greater or equal to T, where

$0 < T \leq 1$. The overlap degree between two instances (i.e. sub-windows) is the fraction of their overlap area over their union area. To this end, our formulation can be expressed as follows:

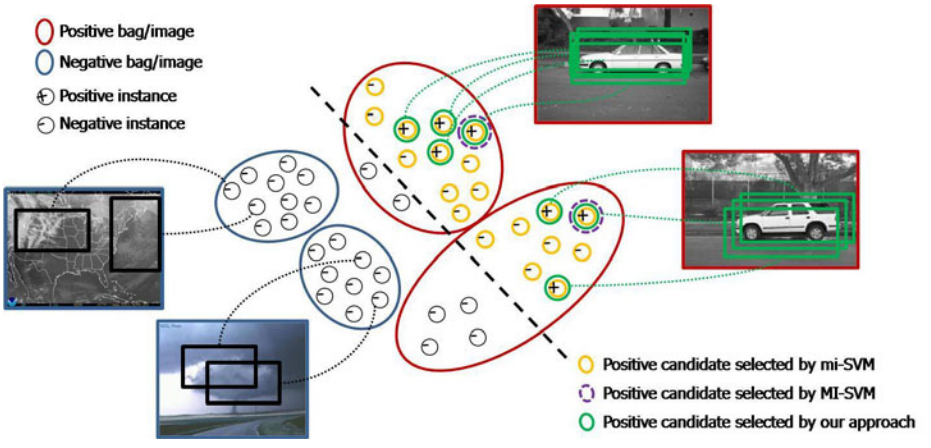

Fig. 1. Illustration of positive candidate selection for the next learning round by different approaches. mi-SVM selects all temporal positive instances (*orange*). MI-SVM selects only the most positive instance per positive bag (*dash-purple*). Meanwhile, our approach selects a subset of spatially related instances (*green*) per positive bag based on their overlap degree with the most positive instance of the bag.

$$\begin{aligned} \min_{\{y_i\}} \min_{\{w,b,\xi\}} \quad & \frac{1}{2}\|w\|^2 + C\sum_I \xi_I \\ \text{subject to} \quad & \forall I:\ Y_I = -1 \wedge -\langle w, x_i\rangle - b \geq 1 - \xi_I\,,\ \forall i \in I\,, \\ & \text{or} \quad Y_I = 1 \wedge \langle w, x^*\rangle + b \geq 1 - \xi_I\,, \\ & \quad \forall x^* \in \mathcal{SR}(x_{mm(I)}, T)\,, 0 < T \leq 1, \text{ and } \xi_I \geq 0 \end{aligned} \tag{4}$$

This formulation can be cast as a mixed integer program in which integer variables are the selectors of $x_{mm(I)}$ and instances in $\mathcal{SR}(x_{mm(I)}, T)$. This problem is hard to solve for the global optimum. However, we exploit the fact that if integer variables are given, the problem reduces to a quadratic programming (QP) that can be solved. Based on that insight, our solution is as follows.

Pseudo code for heuristic algorithm

```
Initialize: for every positive bag B_I
  Compute x_I = Σ_{i∈I} x_i/|I|.
  SR_I = x_I.
REPEAT
```

- Compute QP solution w, b for dataset with positive samples $\{SR_I : Y_I = 1\}$ and negative samples $\{x_i : Y_I = -1\}$.
- Compute outputs $f_i = \langle w, x_i \rangle + b$ for all x_i in positive bags.
- FOR (every positive bag B_I)
  Set $x_I = x_{mm(I)}$, $mm(I) = \arg\max_{i \in I} f_i$
  $SR_I = FindSurround(x_I, T)$
- END

WHILE ($\{mm(I)\}$ have changed)
OUTPUT (w, b)

In our pseudo code, $FindSurround(x_I, T)$ is the function to find instances (i.e. sub-windows) surrounding x_I and have an overlap degree with x_I greater than or equal to T. The greater T is chosen, the fewer instances (i.e. sub-windows) surrounding x_I are selected. Thus, T can be considered as a trade-off parameter for expanding the search space as well. T is a predefined number and is fixed throughout learning iteration. The optimal T is obtained automatically by cross validating on the training set. Additionally, negative candidates of all learning rounds are instances of the negative bags.

5 Experiments

5.1 Dataset

We perform experiments on Caltech benchmark datasets.

- **Caltech 4** contains images of 4 object categories: airplanes (1,075 images), cars_brad (1,155 images), faces (451 images), motorbikes (827 images), and a set of 900 clutter background images.
- **Caltech 101** consists of images in 101 object categories and a set of clutter background images [8]. Each object category contains about 40 to 800 images.

Ground-truth annotations indicating object's locations in images are available for all object categories (but cars_brad category in Caltech 4). These are challenging datasets because of their large variations in object appearance and background. Some example images are shown in Figure 2.

We evaluate the performance of the approaches on binary categorization tasks which are distinguishing images of each object category from background images. On the Caltech 101 dataset, with each binary classification task, a set of 15 positive images taken from one object category and 15 negative images from the background category are given for training; 30 other images from both categories are used for testing. The correlative numbers of positive images, negative images and testing images on Caltech 4 dataset are 100, 100 and 200 respectively. All images are randomly selected.

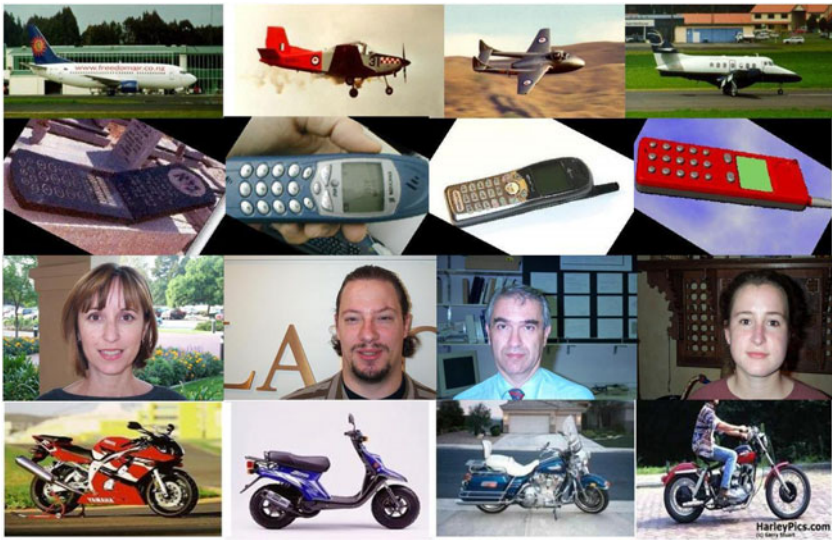

Fig. 2. Example images taken from Caltech 101. From top to bottom are images of airplanes, cellphones, faces and motorbikes respectively.

5.2 Bag and Instance Representation

In order to apply the MIL approaches, we treat bags as images and instances of a bag as sub-windows in the image. We employ the standard Bag-of-Word (BoW) approach for feature representation. First, on each image, we sample a set of points using a grid. The sampling grid has an 8-pixel distance between adjacent points. Then, we use the SIFT descriptor to extract SIFT feature at each point. The SIFT descriptor frame has a 16-pixel width. All descriptors are then quantized using a visual codebook with 100 visual words obtained by applying K-Means to 100,000 training descriptors. Finally, the sub-windows of the image are represented by using a histogram of visual words appearing inside the sub-window region.

5.3 Evaluated Approaches

We compare our approach with the original SVM-based MIL approaches - mi-SVM and MI-SVM - and two other standard approaches called GH and MA. GH denotes a traditional approach in which SVM is used to classify images represented by a histogram of visual words on the whole image region (GH stands for Global Histogram). Meanwhile, MA is an approach that uses tight object rectangles given manually as positive examples and a set of randomly selected windows from negative images - ten windows per negative image - as negative examples

for training (MA stands for Manual Annotation). The measure for comparison is the accuracy ratio with respect to image classification performance. To obtain the best performance of the approaches for fairness, all parameters are optimized. Kernel parameters for SVM and overlap threshold T of our approach are automatically obtained by using the grid-search approach together with 5-fold cross validation.

5.4 Experimental Results

Table 1 and Table 2 list the classification performances of the approaches on Caltech 4 and Caltech 101. Our proposed approach is superior to the others in most object classes. This means the most discriminative instances found by our approach are more meaningful than the one selected by MI-SVM and is also more discriminative than the object regions classified by MA. Moreover, these results prove that our arguments on the effectiveness of using the spatial relation and the limitations of the instance selection criteria of mi/MI-SVM are valid. Because of adding all possible positive instances, mi-SVM also adds more noise to learning and its performance consequently suffers. MI-SVM has a better accuracy than mi-SVM, but it is still worse than ours because of its limited search space.

Table 1. Average classification accuracy of the evaluated approaches on Caltech 4. MA: trains SVM using manual annotation of object region in images. GH: trains SVM using global histogram of images. mi/MI-SVM: MIL approaches proposed by Andrews et al [5]. Note that the performance of MA is computed on 3 categories (airplanes, faces and motorbikes) due to the lack of ground-truth object box of the category cars_brad.

Approaches	Average Classification Rate(%)
MA	90.73
GH	94.46
mi-SVM	72.54
MI-SVM	95.74
Ours	**96.28**

Table 2. Average classification accuracy of the evaluated approaches on Caltech 101

Approaches	Average Classification Rate(%)
MA	78.32
GH	83.37
mi-SVM	60.49
MI-SVM	84.25
Ours	**86.89**

Table 3. Average classification accuracy of the evaluated approaches on 10 categories of Caltech 101

	MA	GH	mi-SVM	MI-SVM	Ours
Butterfly	76.7	76.7	53.3	86.7	**93.3**
Camera	70.0	80.0	53.3	73.3	**86.7**
Ceiling_fan	70.0	**80.0**	53.3	66.7	**80.0**
Cellphone	80.0	**90.0**	63.3	83.3	**90.0**
Laptop	80.0	76.7	66.7	76.7	**86.7**
Motorbikes	73.3	**93.3**	63.3	80.0	90.0
Platypus	83.3	90.0	53.3	86.7	**100.0**
Pyramid	**90.0**	**90.0**	63.3	76.7	**90.0**
Tick	76.7	83.3	56.7	80.0	**90.0**
Watch	**80.0**	**80.0**	53.3	73.3	**80.0**

6 Conclusion and Future Work

We proposed an extension of the SVM-based Multiple Instance Learning framework for image categorization by integrating spatial relations between instances into the learning process. Experimental results on the benchmark dataset show that our approach outperforms state-of-the-art SVM-based MIL approaches as well as standard categorization approaches. To the best of our knowledge, this is the first MIL approach that considers sub-window overlapping relations on image space rather than feature space only. For future work, we want to extend our MIL framework so it can be applied to weakly supervised object localization and recognition.

References

1. Dietterich, T., Lathrop, R., Lozano-Perez, T.: Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence, 31–71 (1997)
2. Maronand, O., Lozano-Perez, T.: A Framework for Multiple Instance Learning. In: Advances in Neural Information Processing Systems, pp. 570–576 (1998)
3. Zhang, Q., Goldman, S.: EM-DD: An Improved Multiple Instance Learning Technique. In: Advances in Neural Information Processing Systems, pp. 1073–1080 (2002)
4. Chen, Y., Wang, J.Z.: Image Categorization by Learning and Reasoning with Regions. Journal of Machine Learning Research, 913–939 (2004)
5. Andrews, S., Tsochantaridi, I., Hofmann, T.: Support Vector Machines for Multiple-Instance Learning. In: Advances in Neural Information Processing Systems, pp. 561–568 (2003)
6. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-Instance Learning via Embedded Instance Selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1931–1947 (2006)
7. Fu, Z., Robles-Kelly, A.: An Instance Selection Approach to Multiple Instance Learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 911–918 (2009)

8. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Workshop on Generative-Model Based Vision, IEEE Conference on Computer Vision and Pattern Recognition (2004)
9. Nguyen, M.H., Torresani, L., Torre, F., Rother, C.: Weakly Supervised Discriminative Localization and Classification: A Joint Learning Process. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
10. Galleguillos, C., Belongie, S.: Context Based Object Categorization: A Critical Survey. In: Computer Vision and Image Understanding (2010)
11. Marques, O., Barenholtz, E., Charvillat, V.: Context Modeling in Computer Vision: Techniques, Implications, and Applications. Journal of Multimedia Tools and Applications (2010)
12. Zha, Z.J., Hua, X.S., Mei, T., Wang, J., Qi, G.J., Wang, Z.: Joint Multi-Label Multi-Instance Learning for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
13. Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A., Hebert, M.: An Empirical Study of Context in Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1271–1278 (2009)
14. Wolf, L., Bileschi, S.: A Critical View of Context. International Journal of Computer Vision (2006)