Attention Region Latent SVM for Image Classification

Shengan Zhou [1], Peng Liang[2,b] and Jiangwei Qin[3]

[1] *School of Electronic Information System, Guangdong Vocational Institute of Public Administration, GuangZhou, 510800, China*

[2] *School of Computer Science, GuangDong Polytechnic Normal University, Guangzhou, 510665, China*

[3] *College of Medical Information Engineering, GuangDong Pharmaceutical University, Guangzhou, 510006, China*

[b]*cs_phoenix_liang@163.com*

Abstract

This paper presents a new method for image classification based on image saliency region. The proposed attention region latent SVM (ARLSVM) is highly distinctive by training in a weakly-supervised manner which without requiring objects position or bounding boxes in training images. We use a latent SVM to model the optimization problem with saliency regions are latent variables. An EM method is proposed to solve the semi-convex optimization problem. Through experiments, our proposed approach performs favourably compared with two well-known algorithms in a benchmark dataset.

Keywords: saliency map, SVM, image classification, optimization problem.

1. Introduction

Recent researches proved that the performance of object classification can be highly improved by giving bounding boxes around exemplars of a given class label. Estimating the distribution of object location i.e. the image saliency is an important problem in computer vision [1, 2, 3, 4]. Here we show that the classification of object classes without their localization, also benefits from the estimation of image saliency map, even when these are not supplied as part of the training. What's more, we find that given a class, visual saliency is attributed to different local regions based on their appearance and their spatial location in an image i.e. the human faces also be the middle part of the image.

Visual saliency local regions have been detected by using interest points (e.g. [5, 6]) which can be invariant to image transformations (e.g. rotation, scale, affine). They have been very successfully applied in image classification, object recognition and object detection. [7, 8]. Moosmann et al. [9] use image saliency maps as a supervised information to improve object categorization. Xx segment the image into a number of regions, and associate image labels with these segmented regions for object classification. Gao and Vasconscelos [10] define a joint distribution over visual features extracted from the regions and image labels. Parikh et al. [11] explore the use of context to determine which low-level appearance cues in an image are salient or representative of an image's contents. Khan et al. [12] model color based saliency by splitting the cues in a bottom-up descriptor cue and a top-down attention cue. Yao et al. [13] explore fine image statistics and identify the discriminative image patches for recognition.

Recently, latent support vector machine (LSVM) classifiers have shown impressive performance in many visual tasks. Felzenszwalb et al. [14] combine a margin-sensitive approach for data-mining hard negative examples with a latent SVM. Bilen et al. [15] use LSVM model for image classification with object position and size as latent variables.

In this paper we propose a new image saliency based method for image classification. We first divide images into regions, and we associate a binary latent variable with each region indicating whether we include it for "figure" or "ground". The image score is obtained by computing maximum score of each latent variable. We use a proposed ARLSVM to model the optimization problem in a weakly-supervised manner. Finally, an EM method is proposed to solve the semi-convex optimization problem.

2. Attention-Region latent SVM

2.1 Feature selection

Our model is based on learning weights for image attention regions by solving a max-margin problem. The image saliency maps are used as latent variables which cannot be observed either in training images or test images. In our model, the latent variable can be interpreted as "figure" or"ground" in an image. For example in a binary object classification task, the learning weights are used as a score function to classify object and the context in which it appears.

We segment an image into a set of N regions $\{x_1, x_2, ..., x_N\}$ by using a user specified feature method such as spatial pyramid with uniform grid[17] or normalized cuts [23]. The image region is represented by a bag-of-word histogram over quantized SIFT descriptors [16]. We define saliency map of an image as a set of saliency region $Z = \{z_1, z_2, ..., z_N\}$. Each saliency region is a latent variable. The value of $z_i \in \{1, ..., K\}$ indicates how the segment will be scored.

2.2 Image score function

The image score function $f(x,z)$ is a discriminative function that measures the matching quality between given image features x and predicted class label y with latent variable z . The separating hyperplane can be maximized by changing the saliency map of image.

Given a particular assignment of latent variable, image score function is defined as inner-class and intra-class terms. Inner-class term is represented as the sum of unary and pairwise, while intra-class term is defined as a score punishment proportional to the deviation of the average image saliency map $\bar{x}$:

$$f(x,z)=\sum_{k=1}^{K}\sum_{i=1}^{N}w_k^T x_i+\sum_{k,l=1}^{K}\lambda_{kl}\sum_{(i,j)\in E}p(x_i,x_j)+\sum_{k=1}^{K}\delta_k\sum_{i=1}^{N}p(x_i,\bar{x}_i) \tag{1}$$

The unary score is determined by the given value of latent variable z_i. The pairwise term denotes that image regions with the same value of latent variables usually connected spatially contiguous. The intra-class term is inspired by the fact that the corresponding local feature pairs will fall within the same area in geometric invariant space, which can be showed in Fig. 1.

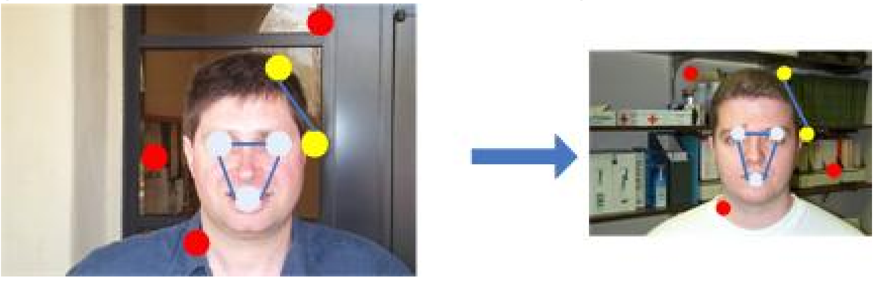

Fig.1: Corresponding pairwise local features between two images

Then the score function of an image is defined as maximum over the scores obtained over all possible values of the latent variables $f(x)=\max_{z} f(x,z)$

2.3 Problem formulation and optimization solving

We use the latent SVM framework described in [1, 9, 17] to learn the model weights vector z with the image saliency maps s are latent variables. Given an image set $m\in M$ with class labels $y^m\in\{1,-1\}$, the optimization with hinge loss is represented as followed:

$$L=\min_{w}\frac{1}{2}\|w\|^2+C\sum_{m=1}^{M}\max(0,1-y^m f(x^m,z)) \tag{2}$$

Where $f(x^m, z)$ is score function of image m. While using ξ_i to instead of $\max(0, 1 - y^m f(x^m, z))$, the optimization formula can be equivalently rewritten as:

$$\begin{aligned} &\min_{w,\xi_i \geq 0} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{\Omega} \xi_i \\ s.t. \quad &\exists z \ \ f(x^m, z) \geq 1 - \xi_m, \ if \ y^m = 1 \\ &\forall z \ \ f(x^m, z) \leq -1 + \xi_m, \ if \ y^m = -1 \end{aligned} \tag{3}$$

According to Eq. 3, the optimization problem is semi-convex in the sense that the objective function is convex if the latent variables for positive images are fixed. For positive images, a positive prediction required at least one latent variable assignment result in positive label, while all possible assignment of latent variables should result in negative prediction.

We use EM algorithm to solve the problem with w and z as two blocks of variables, which is alternately optimized while keeping the other variable fixed. The detail of EM Algorithm is shown in Table 1 as followed:

Table 1 The proposed EM algorithm

```
While $t = 1, 2, ..., T$ do
    Choose a random training image $m$
    Fix parameter $w$, Calculate $z^* = \max_z f(x, z)$
    Use $f(x^m, z^*) \geq 1 - \xi_m, \ if \ y^m = 1$ take the place of Eq.().
    Fix parameter $z^*$, Calculate $\nabla_w L = w + C \sum_i g_w(x_i)$,
    $g_w(x_m) = \begin{cases} 0 & if\, y_m f(x_m, z) \geq 1 \\ -y_m x_m & otherwise \end{cases}$
end while
```

3. Experiment and Analysis

3.1 Experimental setup

In this section, a well-known image datasets are used in our experiments for evaluating the proposed algorithm. We also compare the proposed algorithm against two well-known methods [19, 20]. We present experiments on airship, camel, cannon, bus, face, car, chair, dog, horse and orangutan from Caltech-256.

The dataset is very difficult to classify due to significant background clutter and large intra-class variability. Some example images of Caltech-256 are shown in Fig.2.

We use average precision (AP) to evaluate the classification performances. AP is computed from the average value of precision over the interval by decreasing classification score. The mean average precision is the mean of the AP computed for each class.

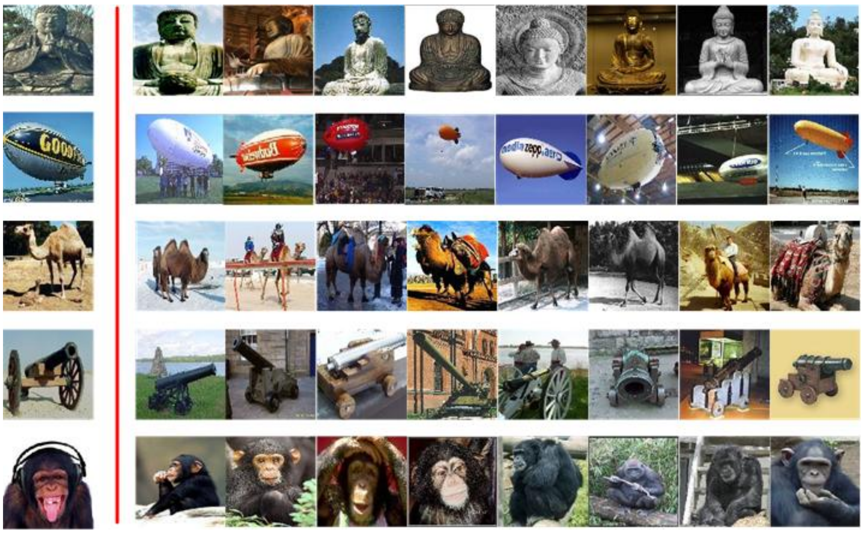

Fig 2. Example images on Caltech-256 dataset

We first select features from image dataset by using Bag-of-Words representation via SIFT descriptor. The SIFT descriptor in this experiment is 16×16 pixel cells.

computed over a grid with 8 pixels. We also perform k-means clustering of a random subset of descriptors from the training set to form a visual vocabulary, and the typical vocabulary size in our experiments is $W = 800$. All the experiments are repeated ten times with randomly selected training set and testing set.

We use the χ^2 -distance to compute the distance between histogram $x_1 = (h_1, ..., h_s)$ and $x_2 = (w_1, ..., w_s)$ shown as followed:

$$d(x_1, \ x_2) = \frac{1}{2} \sum_{i=1}^{s} \frac{(h_i - w_i)^2}{h_i + w_i} \tag{4}$$

3.2 Experimental result analysis

The classification experiment is done by using some categories in Caltech-256 dataset with different algorithms. Each category has 100 to 200 images, and average image size is 300*200 pixels. We use 50 images per class for training classifiers and the rest images for testing.

Table 1. shows the detailed results of classification experiments for three compared algorithm. The best performance of each image category is emphasized by underline. Although the improvement is quite small, the proposed ARLSVM is out performance than the other algorithms for most of the test cases. More important, categories of vehicles appearing in urban street scenes (bus and car) standout by particularly large improvements for both proposed ARLSVM and RBLSVM. This can be explained by these categories have more complex background, which are easily confused by SVM via BoW histogram.

It is interesting to comparing the proposed ARLSVM and RBLSVM on with face, orangutan, horse, cat and dog categories. Our method achieves better performance in all categories, for this reason that these objects have more stable structure such as human face and horse walking by four legs.

Table 1. Average precision (AP) for bag-of-words SVM, region-based latent SVM and proposed attention-region latent SVM .

Class	BoW-SVM	RBLSVM	ARLSVM
Airship	42.6%	44.0%	45.5%
Camel	38.0%	41.2%	42.0%
Cannon	54.0%	53.3%	52.0%
Bus	59.3%	72.0%	74.6%
Face	80.6%	86.0%	86.6%
Car	68.6%	76.0%	79.2%
Chair	64.6%	71.4%	72.6%
Dog	74.2%	76.6%	71.4%
Horse	60.0%	66.0%	69.0%
Orangutan	70.0%	81.0%	84.0%

4. Conclusion

In this paper, we have presented a new method for image classification based on image saliency region. The proposed ARLSVM is trained in a weakly-supervised manner which without requiring objects position or bounding boxes in training images. We use a latent SVM to model the optimization problem with saliency regions are latent variables. An EM method is proposed to solve the semi-convex optimization problem. Comparing with the bag-of-words SVM, region-based latent SVM and our attention region latent SVM has shown promising results in a benchmark dataset classification experiment, which contain significant clutter and occlusions (Caltech-256). The ARLSVM achieves better performance than other methods.

Despite the advantage described above, our method is not meant as a sufficient solution to all kinds of image classification. In our future work, we will extend our approach to using non-linear kernels, which are known to yield better classification performance.

Acknowledgement

The research work was supported by National Natural Science Funds for Young Scholar of China under Grant No.61003270.

References

[1] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence*, 20(11), pp. 1254-1259, 1998.

[2] C. Koch and S. Ullman. Shifts in selective visual attention: Towards underlying neural circuitry. *Matters of Intelligence*, 188, pp. 115–141, 1987.

[3] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Saliency estimation using a non-parametric low level vision model. In CVPR, pp. 433-440, 2011.

[4] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley. Image saliency: From intrinsic to extrinsic context. In CVPR, pp. 417-424, 2011.

[5] D. Lowe. Distinctive image features form scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2), pp. 91–110, 2004.

[6] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Intl. Journal of Computer Vision*, 60(1), pp. 63–86, 2004.

[7] D. Gao and N. Vasconcelos. Integrated learning of saliency, complex features and object detectors from cluttered scenes. In CVPR, 2, pp. 282-287, 2005.

[8] A. Vedaldi and A. Zisserman. Efficient additive kernels using explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 34(3), pp. 480-492, 2010.

[9] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. In ECCV Workshops, pp. 1-14, 2006.

[10] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition form cluttered scenes. In NIPS, pp. 1-11, 2004.

[11] D. Parikh, L. Zitnick, and T. Chen. Determining patch saliency using low-level context. In ECCV, 5303, pp. 445-459, 2008.

[12] F. S. Khan, J. van deWeijer, and M. Vanrell. Top-down color attention for object recognition. In ICCV, pp. 979-986, 2009.

[13] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In CVPR, pp. 1577-1584, 2011.

[14] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *Pattern Analysis and Machine Intelligence*, 32(9), pp.1627–1645, 2010. 2, 3, 4.

[15] H. Bilen, V. Namboodiri, and L. Van Gool. Object and action classification with latent variables. In BMVC, pp. 1-11, 2011.