Gaze Latent Support Vector Machine for Image Classification Improved by Weakly Supervised Region Selection

Xin Wang, Nicolas Thome, Matthieu Cord

Gaze Latent Support Vector Machine for Image Classification Improved by Weakly Supervised Region Selection

Xin Wang*, Nicolas Thome, Matthieu Cord

Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France

Abstract

This paper deals with Weakly Supervised Learning (WSL), *i.e.* performing image classification by leveraging local information with models trained from global image labels. We propose a new WSL method which incorporates gaze features collected by an eye-tracker to guide the region selection policy. Our approach presents two appealing advantages: gaze features are cheap to collect, *e.g.* one order of magnitude faster than bounding boxes, and our method only requires gaze features during training, while being gaze free at test time. For this purpose, the training objective is enriched with a gaze loss, from which we derive a concave-convex upper bound, leading to an off-the-shelf CCCP optimization scheme. Extensive experiments are conducted to validate the effectiveness of the approach for WSL image classification on two public datasets with gaze annotation, *i.e.* PASCAL VOC 2012 action and POET. In addition, we publicly release a new food-related dataset, the Gaze-based UPMC Food dataset (UPMC-G20), which covers 20 food categories and 2,000 images. This dataset intends to promote the research in the food-related computer vision community.

Keywords: weakly supervised learning, human gaze, image classification

*Corresponding author
Email address: xin.wang@lip6.fr (Xin Wang)

1. Introduction

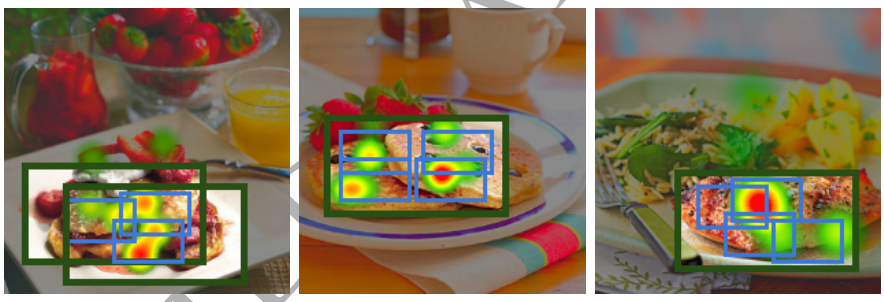
Deep learning has achieved great success in the era of big data. The dominance of deep models is witnessed in the fields of face recognition [1], machine translation [2], speech recognition [3], and even the Go game [4]. One typical example is the success of deep convolutional neural network (DCNN) in computer vision. From the AlexNet [5] to the state-of-the-art deep Residual Networks [6], the DCNN has much outperformed the traditional hand-crafted feature-based machine learning methods, and now being even better than the human experts on the largest classification competition ImageNet Large Scale Visual Recognition Competition (ILSVRC) [7, 6]. Moreover, deep models trained on ImageNet can also be applied effectively to different target domain or different tasks by transfer learning [8]. As a result, state-of-the-art results on standard benchmarks are nowadays obtained with deep features as input. Recent studies show that fine-tuning and data-augmentation can further boost the performance of the transferred models [9].

Although deep models are leading the state-of-the-art of computer vision, current DCNN architectures have limited capabilities in capturing local information in clutter images. However, for the real-world images, local information is critical for identifying the existence of the target object in a clutter environment. Recently, attempts have been made to overcome this limitation by encoding local information by following the design of Bag-of-words (BoW): [10, 11] propose BoW models with deep features as local region activations and [12] developed BoW layers. Despite the encoding contains information locally, the entire image is encoded into the final representation, which may include the irrelevant information. Such clutter information decreases the discriminative power of the model. In this case, expensive annotations such as bounding boxes are often used to localize the target object. Clutter information is subsequently filtered out by omitting the information outside the bounding boxes.

One promising option is to develop weakly supervised learning (WSL) model. Weakly supervised learning (WSL) is a general learning problem which attempts

at making accurate predictions from coarse annotations. For instance, using only image label for segmentation or using the preference relationships between examples for dimensionality reduction problem [13]. In this paper, we focus on applying WSL methods on object classification improved by weakly supervised region selection. Our model learns to remove the clutter background information by selecting the relevant regions without explicit localization information.

Multiple Instance Learning (MIL) [14] is one of the main paradigms for training WSL models. Following this paradigm, an image is described as a bag containing a certain number of instances. Under the hypothesis of MIL, a bag is positive if there is at least one positive instance in the bag, while a bag is negative if all the instances in the bag are negative. Deformable part model (DPM) [15] is a well-known instantiation of MIL, which achieves excellent performance for object detection.



(a) *french toast* (b) *pancakes* (c) *grilled salmon*

Figure 1: Illustration of our model. Multi-scale latent regions are generated in a sliding-window scheme, *e.g.* green (resp. blue) large scale (resp. small scale) regions. Our model can automatically select multiple semantically meaningful regions, *e.g.* those containing the target object class of different scales, from global image labels, *e.g.* french toast. To improve the quality of the region selection, G-LSVM supports regions with gaze information (shown as the heatmap), *e.g.* high density of gaze of positive example for training. (Best viewed in color)

However, selecting relevant regions from images with only image-level annotations is a very challenging task. Intuitively, human gaze is a promising extra information for guiding the search of local semantic information. We

consider here gaze features recorded by an eye-tracker device, which presents two useful properties. One is that gaze features, when collected from people asked to identify a semantic category in an image, contain useful information about the position of the target objects or relevant regions for classification. Another appealing property of gaze features is that they are cheap to collect, especially compared to traditional annotation such as bounding box, typically 1s *vs* 26s [16]. Human gaze carries certain amount of object localization information, but compared with traditional bounding-box annotations, it is weaker for high noise rate and sparse distribution.

Towards a more robust classification model, combining multiple features is a promising choice [17, 18]. In this paper, we propose a new weakly supervised learning model which attempts at incorporating gaze feature collected by an eye-tracker to improve classification performance. Our model, named as G(aze)-LSVM, generalizes latent SVM [15] by exploiting human gaze for localizing objects. Fig. 1 illustrates the rationale of our model for exploiting local information. Latent regions correspond to a set of sub-regions in the image generated by a sliding window scheme. By assigning a scale (size) of the sliding window for each single-scale model, *e.g.* green (resp. blue) large scale (resp. small scale) regions, we then combine a multi-scale model by adapting the object bank representation [19].

G-LSVM is able to automatically select semantically meaningful regions, *e.g.* those containing the target object class of different scales, from global image labels, *e.g.* french toast or pancakes or grilled salmon. To improve the quality of the region selection, G-LSVM supports regions with gaze information, *e.g.* high density of gaze for positive example. Our model is then optimized by reducing a loss function incorporating gaze penalization using the Concave-Convex Procedure (CCCP) [20]. The optimization procedure ensures that our model only needs gaze for training rather for test.

A preliminary version of our model was described in the conference paper [21]. We extend this work at several levels. Firstly, our contributions regarding the WSL model are two-fold: we take into account gaze features for

negative images (Section 3.3) whereas only positive images are used in [21], and also extend the region selection policy from a single region to several regions for performing prediction (Section 3.4), leading to a generalization of top k latent SVM model [22]. Beyond model extension, we also provide a much more thorough experimental analysis for validating the proposed contributions. Last but not least, we publicly release a new food-related dataset with gaze annotation: Gaze-based UPMC Food dataset (UPMC-G20), which will be useful for the community to evaluate gaze models in food oriented applications.

This paper is organized as follows. In section 2, we review gaze analysis and weakly supervised learning methods for computer vision . In section 3 we formally introduce our weakly supervised learning image classification models. In section 4, we introduce the optimization procedure. In section 5, we present our experimental results to validate our models. Specifically, in section 5.2.3, we introduce our new food-related gaze dataset UPMC-G20 in detail. The conclusion is provided in section 6.

2. Related works

2.1. Eye-tracking gaze

Gaze features are appealing since they can reflect the salient parts of the image [23, 24]. People usually use gaze as an extra information channel for solving computer vision problems, such as action recognition [25, 26], object detection [27, 24], saliency prediction [28, 29] and segmentation [30, 16]. In video analysis, since subjects tend to watch at the moving objects, gaze are also widely used to localize important objects [31, 32, 33, 34].

Another appealing property of gaze is that generating gaze by the human is at almost zero-cost. Collecting gaze is more user-friendly and less time-consuming than collecting traditional annotations: it takes about 1 second to collect gaze for one image [16], comparing to 26s for drawing a bounding box [35] and 15-60 min for labeling the segmentation mask for an image [36].

To acquire gaze annotations for different applications, people design various collection protocols [37, 16, 38, 27]. The collection protocols can be grouped into two categories: *task-driven* and *free-viewing*. Task-driven means the annotators are given a specific semantic to look at, *e.g.* a dog. Free-viewing means the annotators view the image freely without specific purpose. As an example of *free-viewing*, Lopez et al. [37] expose simultaneously two images on the screen for evaluating the annotator's visual preference. The aim of this protocol is to collect the gaze features of left and right image for classifying the visual preference. Papadopoulos et al. [16] use an instantiation task-driven protocol. Specifically, this protocol first group image categories into visual-similar pairs. Then the annotation interface exposes to the annotator one image from a selected pair. The annotator should make a decision on the category of the image. The advantage of this protocol is that it does not need the target-absent image to avoid guess, which further reduces unnecessary labeling time. Similarly, Mathe et al. [38] annotate two concepts: *actions* and *context*. One image is exposed to the annotator. Then the annotator is told to find all the actions in the image. Since then, gaze in one image are related to all categories. Gilani et al. [39] use a similar protocol as [38]. But additionally, they have an extra free-viewing protocol for comparing the internal connection with the task-driven protocol. In this paper, we propose a new dataset, UPMC-G20, with gaze annotation using a similar task-driven protocol as in [16]. This dataset is based on the large-scale food-related dataset UPMC Food-101 [40]. The detail of UPMC-G20 is described in section 5.2.3.

2.2. Weakly supervised learning

Multiple Instance Learning (MIL) [14] is one of the main paradigms for training WSL models. Under the assumption of MIL, a bag is positive if there is at least one positive instance in the bag, while a bag is negative if all the instances in the bag are negative. Several attempts have been devoted to applying MIL for object detection[15, 41, 42], scene recognition [43, 44] and dictionary learning [45, 46]. As an instantiation of MI-SVM [47], the latent Support Vec-

tor Machine (LSVM) [15] introduces a theoretically sound formalism for WSL. One challenge with LSVM is due to the introduction of latent variables, which makes the resulting optimization problem non-convex. When using sliding window approaches for generating the candidate regions, the size of the latent space becomes enormous. To overcome this issue, incremental exploration strategies have been proposed in [48, 49]. Finally, recent works focus on enriching the prediction function, by using several (top) instance scores instead of using a single max [22], or by incorporating negative evidence [50, 51, 52].

A growing interest is incorporating gaze information into WSL models. Mathe et al. [53, 54] proposes using reinforcement learning to find a latent space sampling policy from gaze. This method is efficient at the cost of prediction accuracy. Karthikeyan et al. [27] proposes to train a face and text detector from only gaze information. Although this work does not use image features, it still requires bounding boxes to segment out face and text regions. Shcherbatyi et al. [55] integrates gaze into Deformable Part Model for selecting one relevant object location. Their model require gaze annotations for test. Shapovalova et al. [32] and Wang et al. [21] focus on WSL recognition by penalizing region selection with gaze. However, the gaze information is not sufficiently exploited because only positive examples are penalized with gaze. In this paper, comparing to the previous works, our model is generalized to capture the gaze information related to both positive and negative examples. Our model only needs gaze for training rather for test. Since then, our model combines the gaze feature and image feature, but does not use bounding box as input. We also find that the generalization to top k region selection strategy much benefits from the gaze information.

3. G-LSVM: weakly supervised gaze latent SVM

3.1. Latent SVM

We consider the problem of learning with weak supervision in a binary classification context based on the Latent SVM model [15].

The prediction function $f : \mathcal{X} \to \mathcal{Y}$ takes as input an image x, and outputs a binary $y \in \{+1, -1\}$. Each image x is associated with latent variables $z \in Z(x)$, which corresponds to a set of sub-regions. For each region z in image x, we extract a visual feature vector $\Phi(x, z) \in \mathbb{R}^d$, *e.g.* deep features. Our model is linear with respect to Φ, *i.e.* each region z is assigned the score $\langle \mathbf{w}, \Phi(x, z) \rangle$, where $\mathbf{w}$ is learned from data. The problem is weakly supervised since the region-specific labels are unknown during training. Our prediction takes the maximum score over the latent variables:

$$f_{\mathbf{w}}(x) = \max_{z \in Z(x)} \langle \mathbf{w}, \Phi(x, z) \rangle. \tag{1}$$

A standard classification metric is the $0/1$ loss, which means the loss equals $0/1$ if the classification is correct/false. However, $0/1$ loss is difficult to optimize. As in LSVM, we use the hinge loss as a conventional upper-bound of $0/1$ loss. As a result, a classical-SVM like loss is proposed for LSVM:

$$\mathcal{L}_{LSVM}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \max(0, 1 - y_i f_{\mathbf{w}}(x_i)), \tag{2}$$

where y_i is the true label of image x_i, $\hat{y}_i = \operatorname{sgn}(f_{\mathbf{w}}(x_i))$ is the label predicted by our model, hinge loss is defined as $\Delta_c(\hat{y}_i, y_i) = \max(0, 1 - y_i f_{\mathbf{w}}(x_i))$ and $\frac{1}{2}\|\mathbf{w}\|^2$ is the standard max margin regularization term.

3.2. Positive Gaze Latent SVM

This model generalizes latent SVM [15] by biasing the selection of latent regions based on the gaze information during the training scheme. The training objective of G+LSVM is as follows:

$$\mathcal{L}_{G^+}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_c(\hat{y}_i, y_i) + \gamma \cdot \delta_g(\hat{z}_i, x_i, y_i), \tag{3}$$

where z_i is the region with the maximum total duration of fixations, $\hat{z}_i = \arg\max_{z \in Z(x_i)} \langle \mathbf{w}, \Phi(x_i, z) \rangle$ interpreted as the *relevant region* selected by our model. For each training example, Eq. (3) includes a classification hinge loss and a gaze loss δ_g, with a scalar trade-off parameter $\gamma \geq 0$.

The novelty in our training scheme is the introduction of a gaze loss δ_g defined as:

$$\delta_g(\hat{z}_i, x_i, y_i) = \begin{cases} 1 - \frac{\mathrm{g}(x_i, \hat{z}_i)}{\mathrm{g}^+(x_i)} & \text{if } y_i = 1 \\ 0 & \text{if } y_i = -1, \end{cases} \tag{4}$$

where $\mathrm{g}(x_i, z)$ is the density of fixations in the region z for image x_i, $\mathrm{g}^+(x_i)$ is the maximum density of fixations among all the regions of image x_i. Fig. 2 illustrates the proposed gaze loss. In this example, when the color of heatmap is closer to red, the density of gaze is higher. The region contains the maximum density of gaze is shown as z_i (shown as the green rectangle). The gaze loss of z_i is thus defined as 0. The red region z_1 contains a smaller density of gaze with respect to the blue region z_2, leading to a larger gaze loss.

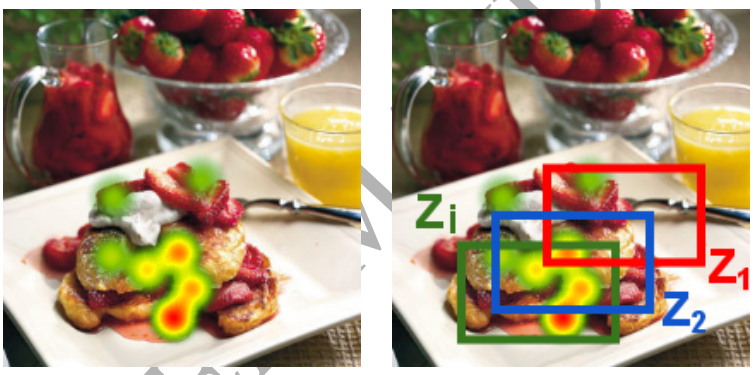


Figure 2: The rationale of the definition of gaze loss. When the color of heatmap is closer to red, the density of gaze is higher. The region contains the maximum density of gaze is shown as z_i (shown as the green rectangle). The gaze loss of z_i is thus defined as 0. The red region z_1 contains a smaller density of gaze with respect to the blue region z_2, leading to a larger gaze loss.

The intuition of training G+LSVM is straightforward. Our training objective in Eq. (3) is biased by the gaze loss δ_g, so that G+LSVM learns a different model parameter $\mathbf{w}$ which tends to minimize gaze loss compared to LSVM. The final decision of our model is to learn a unique $\mathbf{w}$ by compromising between classification loss and gaze loss. In other words, G+LSVM tries to solve the task of classification and localization simultaneously, thus the *relevant region* is

presumed to contain the object of interest, which leads to a better classification result.

Note that given a model parameter $\mathbf{w}$, the *relevant region* $\hat{z}$ only depends on image feature as LSVM, without any gaze information (eq. 1). The benefit of this modeling strategy is that G+LSVM only uses *gaze loss* for training, not for the test. This idea is inspired from *learning using Privileged Information (LUPI)* [56, 57]. The problem addressed by LUPI is that the privileged information is available only at the training stage and is not available at the test stage. By including privileged information into training we obtain a better model, which commits lower generalization error thanks to the localization information for human gaze. This modeling strategy is also practical because models trained with gaze can be applied without gaze annotations. This strategy is also applicable to the following models in section 3.3 and 3.4.

3.3. Positive Negative Latent SVM

One drawback of G+LSVM is the absence of gaze information in negative image. However, a straightforward application of positive gaze loss on the negative image may not work. The reason is that for the positive image, the model should tend to localize where the foreground object is. For the negative image, however, the model should tend to localize where the background is [50]. That's because the overlapping instances between positive and negative example are likely to be the background. According to the *task-driven* protocol, image semantic is related with gaze distribution. Indicated by the gaze, the region with lower density of gaze is more likely to be background. Since then, we should heavily penalize the object region of negative image. This intuition leads to a generalization of G+LSVM, called G±LSVM. In G±LSVM we defined a negative gaze loss, which prefers the region where there is less *objectness*. Contrary to positive image, if a region of negative image contains more gaze, it is force not to be the relevant region of the negative image.

Based on this assumption, we propose a negative gaze loss defined as follows:

$$\delta_g(\hat{z}_i, x_i, y_i) = \begin{cases} 1 - \frac{\mathrm{g}(x_i, \hat{z}_i)}{\mathrm{g}^+(x_i)} & \text{if } y_i = 1 \\ \frac{\mathrm{g}(x, \hat{z}_i) - \mathrm{g}^-(x_i)}{\mathrm{g}^+(x_i) - \mathrm{g}^-(x_i) + \epsilon} & \text{if } y_i = -1 \end{cases} \tag{5}$$

where $\mathrm{g}^-(x_i)$ is the minimum number of gaze among all regions of image x_i, ϵ is set to be 10^{-6}. We subtract the term $\mathrm{g}^-(x_i)$ from the numerator and denominator only to normalize the minimum negative gaze loss to be 0.

We introduce independent parameters γ_+ and γ_- for trading positive gaze loss and negative gaze loss. Assembling all together we get the objective function of G±LSVM:

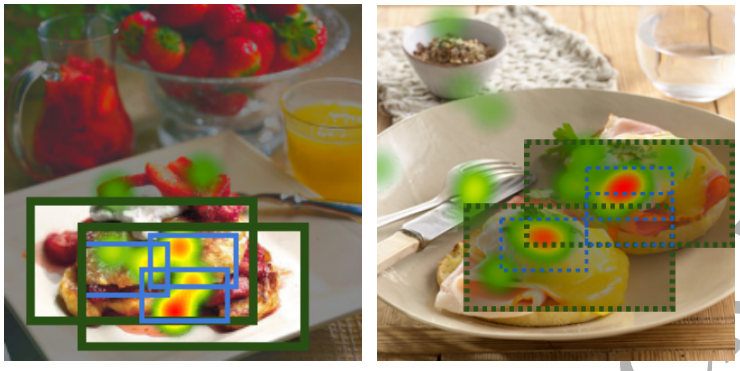
$$\mathcal{L}_{G^+_-}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_c(\hat{y}_i, y_i) + (\llbracket y_i = 1 \rrbracket \gamma_+ + \llbracket y_i = -1 \rrbracket \gamma_-) \cdot \delta_g(\hat{z}_i, x_i, y_i) \tag{6}$$

3.4. Top k instance model

Taking only the maximum scored region as the representative is rigid because one region may be to small to fit an object. To soften the constraint, [22] proposes the definition of soft bags of top k instances. In soft bags, example is represented by the average feature of the top k instances. This method is proved to be robust to the noise in the examples and generalized better than LSVM.

An useful property of top k related to gaze information is its smooth functionality for *sparse gaze limitation*. This limitation is due to the truth that gaze on an image often focus on a small part of the image. For a given example, the gaze loss term has no difference on regions with the same gaze loss. Selection among these regions is random in previous single instance models. This randomness can be eliminated by taking them all via top k strategy.

Fig. 3 illustrates the rationale of our final model. Remind that the goal is to select semantically meaningful regions, *e.g.* those containing the target object class (*eggs benedict* region or its sub-regions in Fig. 3a). By assuming that gaze features are related to regions relevant for the recognition task, gaze and object are matched for positive example. For negative example, top k G±LSVM further



(a) *french toast*, positive example (b) *eggs benedict*, negative example

Figure 3: Illustration of top k G±LSVM model. Human gaze density is represented by the heat map. In our models, positive example emphasize the latent regions with high gaze density (inside the solid boxes), while negative example emphasizes the regions with low gaze density (outside the dashed boxes). Different colors of regions indicate different scales. For one scale, our model takes multiple highest scored regions as the relevant regions. (Best viewed in color)

supports regions with low density of gaze, by assuming that no gaze features are related to classify negative images. Extending the model to top k instances latent SVM can further improve the quality of region selection and reduce the effect of the sparseness of gaze.

The objective function of top k G±LSVM is as follows:

$$\mathcal{L}_{kG^+_-}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_c(\hat{y}_i, y_i) + ([\![y_i = 1]\!]\gamma_+ + [\![y_i = -1]\!]\gamma_-) \cdot \delta_{\mathbf{g}}(\hat{\mathbf{z}}_i, x_i, y_i) \tag{7}$$

where

$$\Delta_c(\hat{y}_i, y_i) = \max(0, 1 - y_i f_{\mathbf{w}}(x_i))$$

$$\delta_{\mathbf{g}}(\hat{\mathbf{z}}_i, x_i, y_i) = \frac{1}{k}\sum_{j=1}^{k} \delta_g(\hat{z}_{ij}, x_i, y_i)$$

$$\hat{\mathbf{z}}_i = \underset{\mathbf{z}\in\mathbf{Z}(x_i)}{\arg\max}\ \langle \mathbf{w}, \Phi(x_i, \mathbf{z})\rangle,$$

where $\mathbf{z}$ is a vector of latent variables, $\mathbf{Z}(x_i)$ the hypothesis space $\{0,1\}^k \setminus \{\mathbf{0}\}$. $\Phi(x_i, \mathbf{z}) = \frac{1}{k}\sum_{j=1}^{k} \Phi(x_i, z_{ij})$.

4. G-LSVM Optimization

In the section 3 we have revised the baseline LSVM, G+LSVM, and proposed three variations, G±LSVM, top k G+LSVM, top k G±LSVM. Each of the five models has a different objective function to optimize. However, notice that when $k = 1$, top k models reduce to the single instance model. Furthermore, when $\gamma_- = 0$, the objective function of G±LSVM (Eq. 6) reduces to G+LSVM (eq. 3), and when $\gamma_+ = 0$, G+LSVM reduces to LSVM (eq. 2). For the reason above, without losing the generality, we only explain how to optimize the most generalized top k G±LSVM objective function Eq. 7 in this section.

Eq. 7 is a sum of hinge loss (classification loss) and gaze loss. We first derive a concave-convex upper bound of Eq. 7 by following steps:

1) *Classification loss part:* For negative example, $y_i = -1$. The second term $1 - y_i f_{\mathbf{w}}(x_i)$ in its classification loss is convex because it is a sum of a constant and a maximum over a set of convex functions. As a result, the sum of the classification loss of all negative examples are convex. For positive example, since $y_i = 1$, it is not convex. We propose to optimize by decomposing the hinge loss of positive example into a difference of two convex functions by applying the following theorem:

$$\max(0, u - v) = \max(u, v) - v, \tag{8}$$

where u, v are two convex functions. The non-convex classification loss of every positive example is thus decomposed as:

$$\max(0, 1 - f_{\mathbf{w}}(x)) = \max(1, f_{\mathbf{w}}(x)) - f_{\mathbf{w}}(x). \tag{9}$$

The maximum of a set of linear functions is convex, so Eq. 9 is a difference of two convex functions.

2) *Gaze loss part:* $\delta_{\mathbf{g}}(\hat{\mathbf{z}}_i, x_i, y_i)$ is difficult to optimize, because the dependency on $\mathbf{w}$ is complex and non-smooth. To overcome this issue, we derive a

convex upper-bound Δ_g, inspired from *margin-rescaling* [58]:

$$\begin{aligned}\delta_{\mathbf{g}}(\hat{\mathbf{z}}, x_i, y_i) \leq \quad & \delta_{\mathbf{g}}(\hat{\mathbf{z}}, x_i, y_i) + \mathbf{w}\cdot\Phi(x_i, \hat{\mathbf{z}}) - \mathbf{w}\cdot\Phi(x_i, \mathbf{z_i}) \\ \leq \quad & \max_{\mathbf{z}\in\mathbf{Z}(x_i)} [\delta_{\mathbf{g}}(\mathbf{z}, x_i, y_i) + \mathbf{w}\cdot\Phi(x_i, \mathbf{z})] - \mathbf{w}\cdot\Phi(x_i, \mathbf{z_i}) \\ \coloneqq \quad & \Delta_{\mathbf{g}}(\hat{\mathbf{z}}, x_i, y_i)\end{aligned} \tag{10}$$

where $\max_{\mathbf{z}\in\mathbf{Z}(x_i)}[\delta_{\mathbf{g}}(\mathbf{z}, x_i, y_i) + \mathbf{w}\cdot\Phi(x_i, \mathbf{z})]$ is a max over linear functions, so it is convex. The second term $\mathbf{w}\cdot\Phi(x_i, \mathbf{z_i})$ is linear. As a result, the difference of the two terms is convex.

Aggregating Eq. 9 and Eq. 10 together, the concave-convex upper bound of the objective function of top k G±LSVM is eq. 11:

$$\begin{aligned}\mathcal{L}_{kG^{+}_{-}}(\mathbf{w}) \leq \overline{\mathcal{L}_{kG^{+}_{-}}(\mathbf{w})} = \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\Bigg[\underbrace{\frac{1}{n_n}\sum_{i_n=1}^{n_n}\max(0, 1 + f_{\mathbf{w}}(x_{i_n}))}_{cn(\mathbf{w})} + \\ & \underbrace{\frac{1}{n_p}\sum_{i_p=1}^{n_p}\max(1, f_{\mathbf{w}}(x_{i_p}))}_{cp_1(\mathbf{w})} - \underbrace{\frac{1}{n_p}\sum_{i_p=1}^{n_p} f_w(x_{i_p})}_{cp_2(\mathbf{w})} + \\ & \underbrace{\sum_{i=1}^{n}\left([\![y_i = 1]\!]\frac{\gamma_+}{n_p} + [\![y_i = -1]\!]\frac{\gamma_-}{n_n}\right)\cdot\Delta_{\mathbf{g}}(\hat{\mathbf{z}}, x_i, y_i)}_{g(\mathbf{w})}\Bigg]\end{aligned} \tag{11}$$

where n_p, n_n are respectively number of positive examples and negative examples. The losses of positive example and negative example are also balanced according to their quantity.

For brevity, we rewrite Eq. 11 as $u(\mathbf{w}) - v(\mathbf{w})$, where:

$$u(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C(cp_1(\mathbf{w}) + cn(\mathbf{w}) + g(\mathbf{w})). \tag{12}$$

$$v(\mathbf{w}) = Ccp_2(\mathbf{w}). \tag{13}$$

We then optimize $u(\mathbf{w}) - v(\mathbf{w})$ by CCCP (algo.1). The CCCP algorithm is guaranteed to decrease the objective function at every iteration and to converge

to a local minimum or saddle point [20]. In Algo 1, the line 3 involves linearizing the concave part $-v(\mathbf{w})$. We calculate the supergradient $\mathbf{v}_t$ of $-v(\mathbf{w})$ at the point $\mathbf{w}_t$, where $\mathbf{v}_t = -\sum_{i_p=1}^{n_p} \Phi(x_i, \hat{\mathbf{z}}_i)$. At line 4, the problem becomes convex, we can use any convex optimization tool for solving this problem.

Algorithm 1: Concave-Convex Procedure

Output: $\mathbf{w}^*$

1 Set $t = 0$, stopping criterion ϵ and initialize $\mathbf{w}$ by $\mathbf{w}_0$, $u(\mathbf{w})$ and $v(\mathbf{w})$ are defined as Eq. 12 and Eq. 13.

2 **repeat**

3 Find hyperplane $\mathbf{v}_t$ to linearize $-v(\mathbf{w})$:
$-v(\mathbf{w}) \leq -v(\mathbf{w}_t) + (\mathbf{w} - \mathbf{w}_t) \cdot \mathbf{v}_t$,

4 Solve $\mathbf{w}_{t+1} = \text{argmin}_{\mathbf{w}} u(\mathbf{w}) + \mathbf{w} \cdot \mathbf{v}_t$,

5 Set t = t+1,

6 **until** $[u(\mathbf{w}_t) - v(\mathbf{w}_t)] - [u(\mathbf{w}_{t-1}) - v(\mathbf{w}_{t-1}))] < \epsilon$;

5. Experiments

5.1. Weakly supervised classification setting

In our models, the first step is generating the latent regions. Latent region set corresponds to square image regions extracted with a multi-scale sliding window strategy. Region size vary from 90% to 30% of the whole image area. For a given scale, a window slides from the upper-left to the bottom-right of the image with a step size 10% in both directions. As a result, for each image, the size of sub-region space varies among `{4, 9, 16, 25, 36, 49, 64}`. Each region is described by the deep features extracted from the `FC7` layer of the pre-trained `imagenet-vgg-m-2048` deep model[1], which are subsequently L2-normalized and add a bias term. In this setting, the size of feature and model parameter are fixed as 2049.

[1] http://www.vlfeat.org/matconvnet/pretrained/

For training the multi-scale model, we adapt the object bank representation [19] for our setting. For a given category, we first train the models independently for all 8 scales (including the full image scale). We then form an 8-dimensional vector for each image by the classification scores and train a linear SVM with $C = 10$ as the multi-scale model. Finally, the multi-scale classification score of all categories are averaged to give an mAP to show the overall performance of our models.

5.2. Datasets

We validate our ideas on three datasets, PASCAL VOC Action dataset annotated with gaze (short for Action) [38], PASCAL VOC Object dataset annotated with gaze (short for POET) [16] and our dataset UPMC-G20. Action contains 4588 images, covering 10 categories. POET contains 6131 images, covering 10 categories out of 20 categories of PASCAL VOC Object dataset. The origin of these images is the `train+val` split of PASCAL VOC dataset. Two sample images of POET and Action are shown with gaze annotations in Fig. 4. UPMC-G20 contains 2000 images, covering 20 food categories. The detail of UPMC-G20 is described in section 5.2.3.

5.2.1. Acquisition protocols of POET and Action

The gaze annotations over these datasets are all collected in task-control manners with slight variations.

1. POET uses the *category specific protocol*, which means that each subject has a specific category of object, *e.g.* cat, to look at. Images in POET may have multiple categories. These multiple classes images are annotated with more than one set of annotations. In out tests, for a positive image, we use the corresponding set of annotations, for a negative image, we calculate the fixation duration for each region of each category, then take the maximum fixation duration across the categories as the fixation duration of this region.

2. Action uses the *category group protocol*, which means the subject is required to find a specific group of categories, *i.e.* actions or context. In other words, if a subject is required to find actions, the subject should find all possible actions in the image. The setting of Action is weaker than POET because annotations are only related with a person, not a specific action.

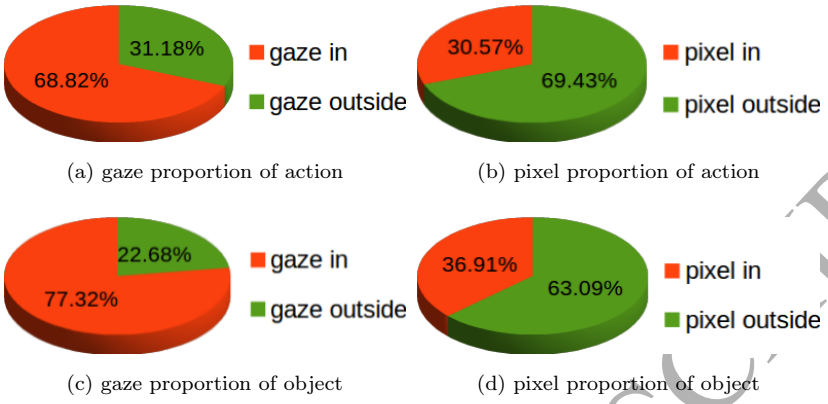
Figure 4: Gaze annotations. *left*: sample image of POET dataset, *right*: sample image of Action dataset. Different colors indicate different observers.

5.2.2. Qualitative analysis of gaze information

We provide a detailed analysis of the gaze data consistency with respect to ground truth bounding boxes on Action and POET. Currently, UPMC-G20 does not have bounding box, so we do not provide the results. We compute statistics for the proportion of gaze falling into or outside of the bounding boxes and compare it to the proportion of image pixels (Fig. 5). Statistically, for action dataset, 68.8% of the gaze fall into the ground-truth bounding-box, while the score of pixels is only 30.6%. Similarly, the scores of object dataset are 77.3% *vs.* 36.9%. This preliminary study provides a quantitative validation that human gaze are highly related to object localization, and convey relevant features for classification.

5.2.3. UPMC-G20 food gaze dataset

UPMC-G20 content. UPMC-G20 is a food-related gaze annotated dataset based on a multi-modal large scale food dataset UPMC-food 101 [40]. We select 20



(a) gaze proportion of action (b) pixel proportion of action

(c) gaze proportion of object (d) pixel proportion of object

Figure 5: Proportions of gaze and pixel numbers in (outside) the ground-truth bounding boxes.

food categories from UPMC-food 101, resulting in 2,000 images. The images selected do not contain text, because it's verified that texts attract attention most [59]. For each image, about 15 fixations across 3 subjects (in average) with a total duration of 2.5 seconds are collected. In total, we have collected 31104 fixations. The categories selected are apple-pie/bread-pudding, beef-carpaccio/beet-salad, chocolate-cake/chocolate-mousse, donuts/beignets, eggs-benedict/croque-madame, gnocchi/shrimp-and-grits, grilled-salmon/pork-chop, lasagna/ravioli, pancakes/french-toast, spaghetti-bolognese/pad-thai.

Samples of images and gaze annotations are shown in the Fig. 6. For full visualization of UPMC-G20, we refer our reader to this page of our dataset: http://webia.lip6.fr/~wangxin/upmcg20/.

Apparatus. Our eye-tracker is a non-invasive Tobii X2-30 with a double eyes gaze sampling rate 30Hz. Eye-tracker is fixed under a 12.6" laptop screen with resolution 1366×768. The subject sits at a distance of about 60cm to the screen. The test environment is quiet and of suitable temperature for not introducing physiological error. The experiment was conducted with the software Tobii Studio (V3.4.5) [60]. Before annotating, for each subject, dominant eye, gender, age are recorded. Before every experiment, Tobii X2-30 is calibrated

Figure 6: Sample images of *apple pie* (left column) and *bread pudding*(right column) in UPMC-G20.

and validated with a standard nine-point procedure to ensure the coordinate of the gaze recorded matches where the subject is looking at. They are taught the procedure of annotation with a clear explanation and validate a simulation test before the formal experiment. Subject record his classification answer by clicking the corresponding option on the screen after viewing an image using a mouse. Comparing to pressing a button to indicate the category as in [16], using the mouse is useful because mousing moving leads to eye moving after every image. The subject then break the possible steady fixating strategy.

UPMC-G20 collection protocol. Our collection protocol is shown in Fig. 7. It is inspired by the *two-alternative forced choice object discrimination* [16]. This protocol is simple to the annotators and can save the time because no irrelevant images for distracting the attention are shown.

The protocol is composed of steps:

1. randomly selecting an image from a pair of categories and exposing for 2.5 seconds, recording the gaze data.

2. making the subject answering a multiple choice question of which the category the image belongs to is shown using a mouse.
3. After exposing every 20 images, a page indicates the progress of the task is shown to heal the anxiety of annotators.
4. After exposing a whole set of images, annotator gets an adequate rest then recalibrate for the next set of images.

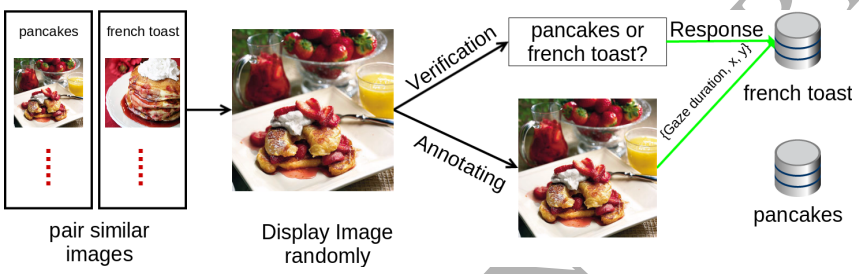


Figure 7: Food gaze collection protocol

5.2.4. Gaze Data

Each gaze is classified into fixation, saccade, or unclassified gaze. For Action and POET, the classification results are already given in the dataset. For UPMC-G20, I-VT filter [61] is used to classified the gaze. Gaze is then represented by fixation in the form of a triplet $(x, y, duration)$. (x, y) is the coordinate of fixation, *duration* is the duration time of this fixation. *Fixation duration* is important since higher exposure time of a fixation reflects a deeper understanding of the local content of the image [62]. The total valid fixation time duration of each subject on each image is normalized to a fixed value. By considering the gaze consistency across subjects, for each region, the fixation duration is summed for all subjects. Gaze loss is calculated for each region using the re-weighted summed fixation.

5.2.5. Dataset split

In order to compare with the state-of-the-art methods, we follow the standard split of `train`, `val`, `test` set as indicated in PASCAL VOC 2012 develop-

ment kit [63]. Since POET contains only 10 out of 20 categories of Pascal VOC 2012 Object, we add back the images of the absent categories in the `train+val` set for training, without gaze information. Finally, our model can be evaluated following the standard protocol. For Action, since by default standard test set requires to identify every person in an image with a bounding box, we conventionally train our model on the training set and test on the validation set. Except for the comparison with the state-of-the-art methods, our experiments are performed by 5 random folds test on the `train+val` set of POET, Action, and the whole dataset of UPMC-G20.

5.3. Comparison with the state-of-the-art

In our model, we set k-G±LSVM with the parameters $C = 10^4$, $\gamma_+ = 0.2$, $\gamma_- = 0.05$ for each scale. In a heuristic manner, for top k models, we set $k = 2, 4, 6, 8$ for scale 90% to 60%, and $k = 10$ for scale 50% to 30%. A multi-scale model is trained as indicated in section 5.1. In all experiments, we use the standard metric mean Average Precision (mAP) as for PASCAL VOC classification.

In Table 1 we show the global score of different methods on the three datasets and the annotations they use. For POET dataset, Deep Fishing [64] and Z&F network [65] are two deep network based methods which only use image label for training. NUS-SCM [66] is an SVM-based method and Oquab [67] is a fine-tuned deep network. They both use training bounding box as the additional annotation. Our method outperforms the four methods with only our weak supervision signals. For Action dataset, we compare with Action part [68] and RMP [69]. The action part is a deep version of poselets and capture parts of the human body under a distinct set of poses, while RMP considers deformation of discriminative parts. They both propose a model with simple annotations (*e.g.* image label and training bounding box) and a model with rich annotations (*e.g.* test bounding box and part annotation. Our model is better than them if they do not use rich annotations. In Table 2 we show the per category performance on the test set of POET. Our model largely outperforms other methods on *boat*,

cat and *diningtable* categories.

5.4. Ablation studies

In this section, we compare LSVM, G+LSVM, G±LSVM and their top k variations. We present the scale-wise classification experiments in Fig. 8. In our model, scale measures the size of the sliding window with respect to the size of the image. In a heuristic manner, for top k models, we set $k = 2, 4, 6, 8$ for scale $90\% - 60\%$, and $k = 10$ for scale $50\% - 30\%$. For most scales, the model performance is better than wSVM (scale=100 in Fig. 8). This result proves the effectiveness of weakly supervised learning: local information is critical for image classification.

We can also observe that adding gaze into the model improve the performance for all scales. The improvement can be explained by two reasons. One is that G+LSVM emphasizes small scales. That is what we expect: for large scales, nearly all regions of positive images are informative, whereas at smaller scales, the model has to focus on relevant localized features. The other is that G±LSVM can also emphasize large scales. Paired T-tests show that G±LSVM is better than LSVM with a larger significance than for G+LSVM, especially for large scale. This phenomenon may have a dual explanation with respect to G+LSVM: not all regions of negative images are non-informative. As a result, for large scale, the ground truth region z_i of negative example has a larger probability to be unique. While for small scales, z_i is selected randomly among all low gaze density regions, which may lead to a less optimal result. When k increases, for small scale, this problem no longer dominates the performance because the set of ground truth regions for negative images is informative with less randomness. We think that is the reason why we constate a substantial performance enhancement at small scales for top G±LSVM.

Table. 3 gives the performance at the smallest scale 30%. At scale 30%, k-G±LSVM (k-G+LSVM) outperform k-LSVM by a margin of 1.8%(1.1%), 1.2%(0.5%), 2.3%(1.2%) for respectively Action, POET and UPMC-G20. Paired T-tests show that k-G±LSVM (k-G+LSVM) is more significant than LSVM

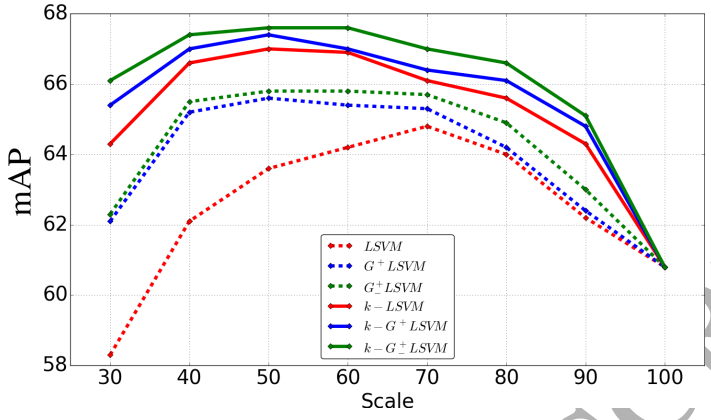
for a risk less than 0.2%(1.0%), 1.0%(2.0%), 0.2%(0.5%) for respectively Action, POET, UPMC-G20. These statistical results show that k-G±LSVM is better than k-G+LSVM with significance at small scale. Top k models much outperform single instance models. Interestingly, as we expected, the gain of k-G±LSVM with respect to k-G+LSVM is much larger than the gain of G±LSVM with respect to G+LSVM.

	Action	POET	label	train BB	test BB	part	gaze
Deep Fishing [64]	-	79.9	√				
Z&F [65]	-	81.2	√				
RMP [69]	65.1	-	√				
NUS-SCM [66]	-	84.3	√	√			
Oquab [67]	-	84.5	√	√			
Action part [68]	64.6	-	√	√			
RMP [69]	71.4	-	√	√	√		
Action part [68]	71.0	-	√	√	√	√	
k-G±LSVM (ours)	69.6	85.9	√				√
G+LSVM [21]	66.8	82.6	√				√
wSVM	59.1	79.8	√				

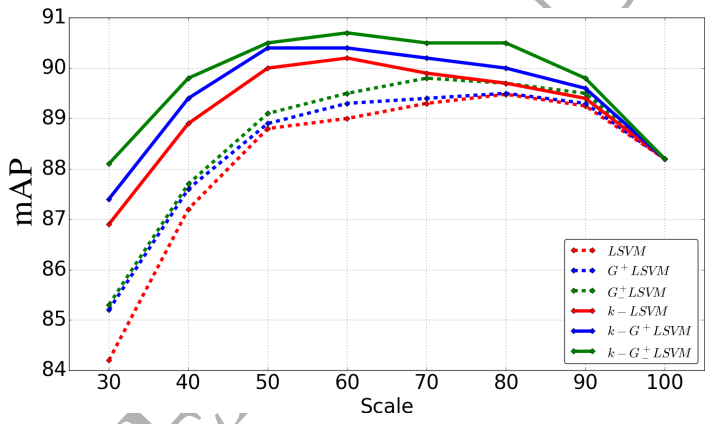
Table 1: Comparison with the state-of-the-art methods on the test set of Pascal VOC 2012 Object, and the validation set of *Action*. Our model outperforms other methods even when they use `global label + training bounding box`. We also achieve comparable results with respect to the models using accurate annotations such as `test bounding box` and/or `human part annotation`.

5.5. Study of hyper-parameters

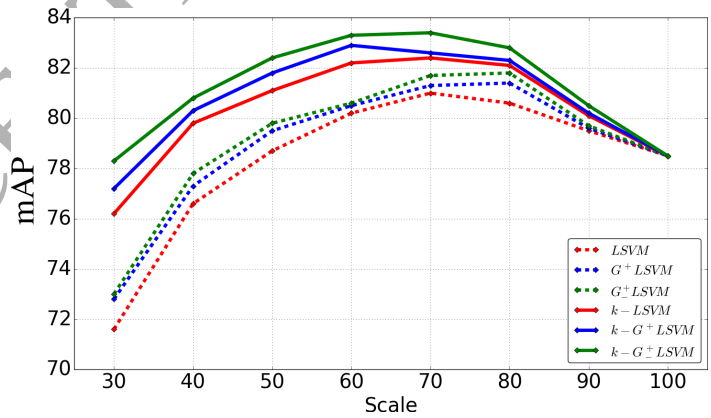
We investigate the impact of the three hyper-parameters in our model: trade-off parameters γ^+, γ^- and k. The impact of the parameter γ_+ of G+LSVM is shown in Fig. 9 for small scale 50%, with k set to be 1. The performances in Fig. 9 are shown on average for all categories. For all three datasets, mAP reaches the peak when γ_+ is in the interval $[0.1, 0.3]$. Note that when γ_+ gets too high, mAP gets even lower than not adding gaze (Fig. 9). Fix γ^- to be the



(a) Action dataset



(b) POET dataset



(c) UPMC-G20 dataset

Figure 8: mAP(%) at different scales. In our model, scale measures the size of the sliding window with respect to the size of the image. Our model outperforms the whole image for most scales using top k instances. Also, k-G±LSVM significantly outperforms other G-LSVM variations at all scales.

POET	mAP	plane	bike	boat	cat	cow	table	dog	horse	motor	sofa
Deep Fishing [64]	79.9	95.0	76.6	82.9	88.6	65.4	69.8	86.5	82.1	85.1	57.0
Z&F [65]	81.2	96.0	77.1	85.5	91.2	74.4	67.7	87.8	86.0	85.1	61.1
NUS-SCM [66]	84.3	**97.3**	**84.2**	85.3	89.3	77.8	75.1	83.0	**87.5**	**90.1**	**73.4**
Oquab [67]	84.5	94.6	82.9	84.1	90.7	**86.8**	69.0	92.1	93.4	88.6	62.3
k-G±LSVM (ours)	**85.9**	97.2	83.9	**90.1**	**94.7**	77.4	**77.3**	**92.3**	87.3	89.9	68.9
G-LSVM [21]	82.6	96.5	80.2	87.7	92.4	71.1	74.1	89.6	84.3	87.5	62.7
wSVM	79.8	95.4	79.6	86.7	92.2	59.6	69.9	90.0	86.7	79.3	58.4

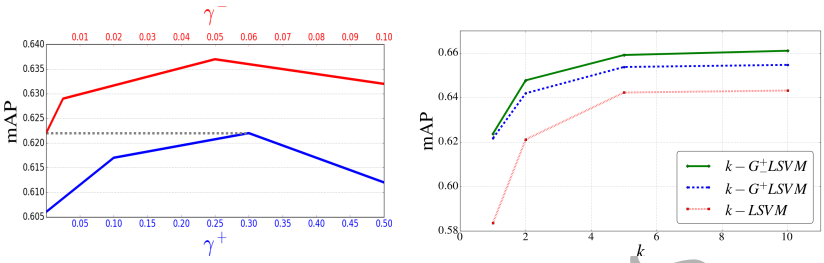
Table 2: mAP(%) per category on the test set of PASCAL VOC 2012 Object.

	Action	POET	UPMC-G20
k-G±LSVM	66.0 ± 0.9	88.1 ± 1.2	78.3 ± 1.0
k-G+LSVM	65.3 ± 1.0	87.4 ± 1.0	77.1 ± 1.1
k-LSVM	64.2 ± 0.8	86.9 ± 1.1	76.0 ± 1.2
G±LSVM	62.4 ± 0.9	85.3 ± 1.1	$73.0. \pm 0.8$
G+LSVM	62.1 ± 0.8	85.2 ± 1.0	72.9 ± 0.9
LSVM	58.2 ± 1.0	84.2 ± 1.1	71.6 ± 1.0

Table 3: mAP(%) of scale 30% on Action, POET and UPMC-G20 datasets. Here we set $k = 10$.

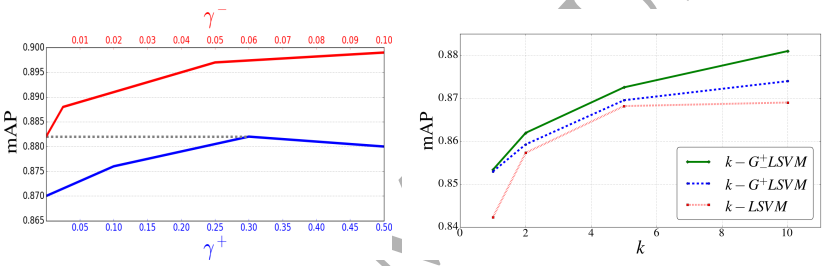
best value obtained by cross-validation, for γ^- , the effective value is found to be a relatively small value between $[0.05, 0.1]$. This result is reasonable because our objective is classification with gaze information as auxiliary information, so the gaze loss should tend to have a smaller weight than the classification loss. The performance of k model varies in the similar trend.

We show in Fig. 9 that our model outperforms k model significantly for all k value at scale 30%. We set γ_+ of G±LSVM and G+LSVM to 0.2, γ_- of G±LSVM to 0.05. From Fig. 9, we also find that by increasing k, gaze latent SVM always outperforms latent SVM. This result signifies that gaze helps better select the regions even when the number of candidate regions largely increases. Heuristically, for selecting k, the small scales prefer a larger k. That's because, for small scale, more regions are semantic for positive images and can smooth



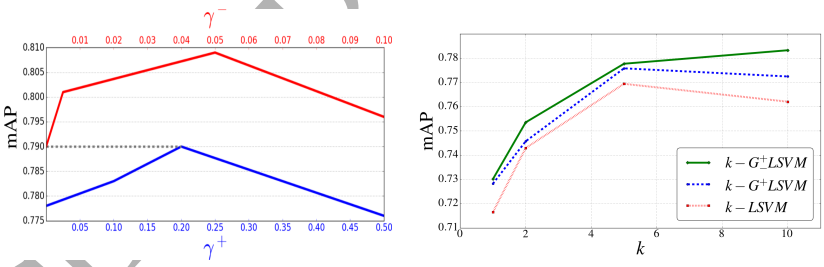
(a) Action dataset

(b) Action dataset



(c) POET dataset

(d) POET dataset



(e) UPMC-G20 dataset

(f) UPMC-G20 dataset

Figure 9: The sensitivity of hyper-parameters γ_+ and k. *left*: At scale 50%, the performance with respect to γ_+ (γ_-) is found to reach the peak value in the interval $[0.1, 0.3]$ ($[0.05, 0.1]$). *right*: At scale 30%, generally, the larger k is, the better the performance is.

the selection of ground-truth regions of negative examples.

5.6. Localization results

The relevant regions proposed by our models are interpretable. We show in Fig. 10 the predicted regions for the model k-G±LSVM at scale 30%, where $k = 10$. We present the first three high scored regions for visual clarity.

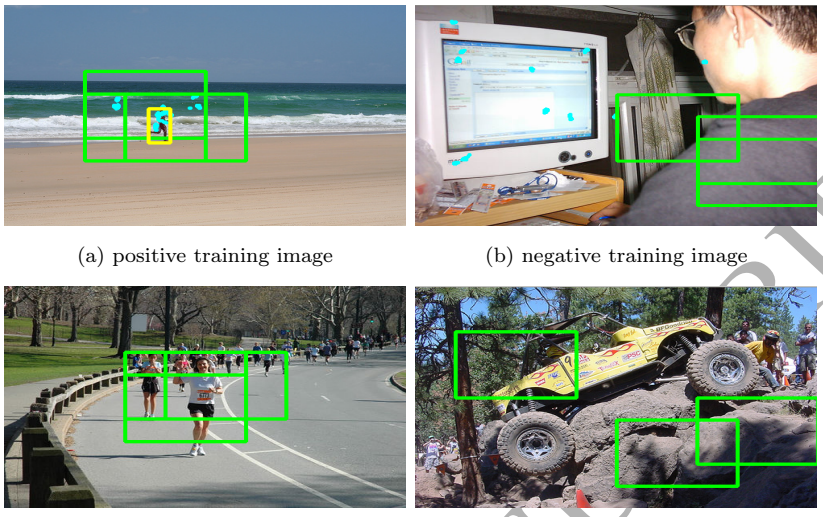
Results for training images are shown in the first row: we show that k-G±LSVM selects areas with more (fewer) gaze for positive (negative) images. Results for test images are shown in the second row, of which gaze features are not available. k-G±LSVM extracts regions which are highly semantic for positive images and extract background for negative images. For example, we find that *running* model has a good result on the positive image. Also for the negative image, the *running* model fires at the regions which have a similar visual semantic to the road and trees. Interestingly, these regions often appear as the background in the *running* images. As these regions have a relatively low density of gaze, our model emphasizes the importance of these regions.

6. Conclusion

In this paper, we introduce G(aze)-LSVM, a weakly supervised learning multi-scale model using sub-region strategy incorporating human gaze for image classification. Our model exploits gaze for guiding the selection of region which is relevant with the image semantic. Furthermore, we find that generalizing the model to the selection of k maximum scored regions can also benefit from the gaze information. Our model leverage human gaze features for training, while the test is gaze free. Experimental results show that our model achieves competitive results with respect to the state-of-the-arts methods on Pascal VOC Action and Object. We also publicly release a medium-sized food dataset with gaze annotation, UPMC-G20, covering 20 categories and 2000 images.

Acknowledgement

This research was supported by the Project VISIIR, ANR-13-CORD-0009.



(a) positive training image (b) negative training image

(c) positive test image (d) negative test image

Figure 10: Localization results achieved by *running model*. (a)(b): training results, (c)(d): test results.

References

References

[1] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: A survey, Advances in Face Detection and Facial Image Analysis (2016) 189–248.

[2] J. Zhou, Y. Cao, X. Wang, P. Li, W. Xu, Deep recurrent models with fast-forward connections for neural machine translation, Transactions of the Association for Computational Linguistics (TACL) 4 (2016) 371–383.

[3] G. Saon, T. Sercu, S. J. Rennie, H. J. Kuo, The IBM 2016 English conversational telephone speech recognition system, in: Interspeech, 2016, pp. 7–11.

[4] Silver David, Huang Aja, et.al, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.

[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[8] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Advances in Neural Information Processing Systems (NIPS), 2014, pp. 3320–3328.

[9] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: British Machine Vision Conference (BMVC), 2014.

[10] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.

[11] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: European Conference on Computer Vision (ECCV), 2014, pp. 392–407.

[12] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, J. Sivic, Netvlad: CNN architecture for weakly supervised place recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5297–5307.

[13] C. Xu, D. Tao, C. Xu, Y. Rui, Large-margin weakly supervised dimensionality reduction, in: International Conference on Machine Learning, 2014, pp. II–865–II–873.

[14] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial Intelligence 89 (1997) 31–71.

[15] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[16] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, V. Ferrari, Training object class detectors from eye tracking data, in: European Conference on Computer Vision (ECCV), 2014, pp. 361–376.

[17] Y. Luo, Y. Wen, D. Tao, J. Gui, C. Xu, Large margin multi-modal multi-task feature extraction for image classification, IEEE Trans. Image Processing 25 (1) (2016) 414–427.

[18] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, Y. Wen, Multiview vector-valued manifold regularization for multilabel image classification, IEEE Transactions on Neural Networks and Learning Systems 24 (5) (2013) 709–722.

[19] L. jia Li, H. Su, L. Fei-fei, E. P. Xing, Object bank: A high-level image representation for scene classification & semantic feature sparsification, in: Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.

[20] A. L. Yuille, A. Rangarajan, The concave-convex procedure (CCCP), in: NIPS, 2001, pp. 1033–1040.

[21] X. Wang, N. Thome, M. Cord, Gaze latent support vector machine for image classification, in: IEEE International Conference on Image Processing (ICIP), 2016, pp. 236–240.

[22] W. Li, N. Vasconcelos, Multiple instance learning for soft bags via top instances, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4277–4285.

[23] S. Ramanathan, V. Yanulevskaya, N. Sebe, Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements, in: International Conference on Multimedia, 2011, pp. 33–42.

[24] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, T. L. Berg, Studying relationships between human gaze, description, and computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 739–746.

[25] S. Mathe, C. Sminchisescu, Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (7) (2015) 1408–1424.

[26] G. Ge, K. Yun, D. Samaras, G. J. Zelinsky, Action classification in still images using human eye movements, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 16–23.

[27] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, B. S. Manjunath, From where and how to what we see, in: International Conference on Computer Vision, 2013, pp. 625–632.

[28] J. Pan, E. Sayrol, X. Giró i Nieto, K. McGuinness, N. E. O'Connor, Shallow and deep convolutional networks for saliency prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 598–606.

[29] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, R. V. Babu, Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5781–5790.

[30] T. Walber, A. Scherp, S. Staab, Can you see it? two novel eye-tracking-based measures for assigning tags to image regions, in: Advances in Multimedia Modeling, International Conference, 2013, pp. 36–46.

[31] S. Karthikeyan, T. Ngo, M. P. Eckstein, B. S. Manjunath, Eye tracking assisted extraction of attentionally important objects from videos, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3241–3250.

[32] N. Shapovalova, M. Raptis, L. Sigal, G. Mori, Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization, in: Advances in Neural Information Processing Systems, 2013, pp. 2409–2417.

[33] D. Damen, T. Leelasawassuk, W. Mayol-Cuevas, You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance, Computer Vision and Image Understanding 149 (2016) 98 – 112.

[34] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, V. Singh, Gaze-enabled egocentric video summarization via constrained submodular maximization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2235–2244.

[35] Hao Su and Jia Deng and Li Fei-Fei, Crowdsourcing Annotations for Visual Object Detection, in: AAAI Workshop, 2012, pp. 1–6.

[36] Pushmeet Kohli and L'ubor Ladický and Philip H.S. Torr, Robust Higher Order Potentials for Enforcing Label Consistency, Int. J. Comput. Vision 82 (3) (2009) 302–324.

[37] S. Lopez, A. Revel, D. Lingrand, F. Precioso, One gaze is worth ten thousand (key-)words, in: IEEE International Conference on Image Processing (ICIP), 2015, pp. 3150–3154.

[38] S. Mathe, C. Sminchisescu, Action from still image dataset and inverse optimal control to learn task specific visual scanpaths, in: Advances in Neural Information Processing Systems, 2013, pp. 1923–1931.

[39] S. O. Gilani, R. Subramanian, Y. Yan, D. Melcher, N. Sebe, S. Winkler, PET: an eye-tracking dataset for animal-centric pascal object classes, in: IEEE International Conference on Multimedia and Expo, 2015, pp. 1–6.

[40] X. Wang, D. Kumar, N. Thome, M. Cord, F. Precioso, Recipe recognition with large multimodal food dataset, in: IEEE International Conference on Multimedia & Expo Workshops, 2015, pp. 1–6.

[41] X. Wang, Z. Zhu, C. Yao, X. Bai, Relaxed multiple-instance SVM with application to object discovery, in: International Conference on Computer Vision (ICCV), 2015, pp. 1224–1232.

[42] W. Shen, X. Bai, Z. Hu, Z. Zhang, Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images, Pattern Recognition 52 (2016) 306–316.

[43] M. Juneja, A. Vedaldi, C. V. Jawahar, A. Zisserman, Blocks that shout: Distinctive parts for scene classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 923–930.

[44] J. Sun, J. Ponce, Learning discriminative part detectors for image classification and cosegmentation, in: International Conference on Computer Vision (ICCV), 2013, pp. 3400–3407.

[45] X. Wang, B. Wang, X. Bai, W. Liu, Z. Tu, Max-margin multiple-instance dictionary learning, in: International Conference on Machine Learning, 2013, pp. 846–854.

[46] A. Shrivastava, V. M. Patel, J. K. Pillai, R. Chellappa, Generalized dictionaries for multiple instance learning, Int. J. Comput. Vision 114 (2-3) (2015) 288–305.

[47] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems (NIPS), 2002, pp. 561–568.

[48] T. Durand, N. Thome, M. Cord, D. Picard, Incremental learning of latent structural SVM for weakly supervised image classification, in: IEEE International Conference on Image Processing, 2014, pp. 4246–4250.

[49] H. Bilen, V. P. Namboodiri, L. J. V. Gool, Object and action classification with latent window parameters, Int. J. Comput. Vision 106 (3) (2014) 237–251.

[50] H. Azizpour, M. Arefiyan, S. N. Parizi, S. Carlsson, Spotlight the negatives: A generalized discriminative latent model, in: British Machine Vision Conference, 2015, pp. 1–11.

[51] T. Durand, N. Thome, M. Cord, MANTRA: minimum maximum latent structural SVM for image classification and ranking, in: International Conference on Computer Vision, 2015, pp. 2713–2721.

[52] T. Durand, N. Thome, M. Cord, WELDON: weakly supervised learning of deep convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4743–4752.

[53] S. Mathe, C. Sminchisescu, Multiple instance reinforcement learning for efficient weakly-supervised detection in images, CoRR abs/1412.0100.

[54] S. Mathe, A. Pirinen, C. Sminchisescu, Reinforcement learning for visual object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2016, pp. 2894–2902.

[55] I. Shcherbatyi, A. Bulling, M. Fritz, GazeDPM: Early integration of gaze information in deformable part models, CoRR abs/1505.05753.

[56] V. Vapnik, R. Izmailov, Learning using privileged information: Similarity control and knowledge transfer, J. Mach. Learn. Res 16 (2015) 2023–2049.

[57] S. You, C. Xu, Y. Wang, C. Xu, D. Tao, Privileged multi-label learning, in: International Joint Conference on Artificial Intelligence (IJCAI), 2017.

[58] T. Joachims, T. Finley, C. J. Yu, Cutting-plane training of structural svms, Machine Learning 77 (1) (2009) 27–59.

[59] M. P. Hsueh-Cheng Wang, The attraction of visual attention to texts in real-world scenes, Journal of Vision 12 (2012) 1–17.

[60] T. AB, Tobii Studio User's Manual Version 3.4.5 (2016).

[61] A. Olsen, The Tobii I-VT Fixation Filter (2012).

[62] L. Fei-Fei, A. Iyer, C. Koch, P. Perona, What do we perceive in a glance of a real-world scene?, Journal of Vision 7 (2007) 1–29.

[63] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, A. Zisserman, The Pascal visual object classes challenge: A retrospective, International Journal of Computer Vision 111 (1) (2015) 98–136.

[64] A. Gordo, A. Gaidon, F. Perronnin, Deep fishing: Gradient features from deep nets, in: British Machine Vision Conference, 2015, pp. 1–12.

[65] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, 2014, pp. 818–833.

[66] Z. Song, Q. Chen, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1585–1592.

[67] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: IEEE CVPR, 2014, pp. 1717–1724.

[68] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2470–2478.

[69] M. Hoai, Regularized max pooling for image categorization., in: British Machine Vision Conference (BMVC), 2014, pp. 1–12.

Xin Wang is a Ph.D. candidate in the Computer Science department LIP6, at UPMC Sorbonne Universities, Paris. He is under the supervision of Nicolas Thome and Matthieu Cord. He received an M.Sc. in Computer Science from the University of Paris-Sud, France in 2014 and a B.Sc. in Optoelectronics from the Huazhong University of Science and Technology (HUST), China in 2012. His research interests include Computer Vision, Machine Learning, Pattern Recognition and Eye-tracking.

Nicolas Thome is an associate professor at UPMC-Paris 6. He received the Ph.D. degree in computer science from the University of Lyon, France in 2007. In 2008, he was a postdoctoral associate at INRETS in Villeneuve d'Ascq, France. His research interests include computer vision and machine learning, including applications for semantic understanding of multimedia data. He is involved in several French (ANR), European and international (Singapore, Brazil) research projects. He is being coordinator of an ANR project on interactive image retrieval in 2013-2017.

Matthieu Cord received the Ph.D. degree in computer science from the UCP, France, before working in the ESAT lab at KUL University, Belgium, and in the ETIS lab, France as Assistant Professor. He joined the Computer Science department LIP6, at UPMC Sorbonne Universities, Paris, in 2006 as full Professor. In 2009, he was nominated at the IUF (French Research Institute) for a 5 years delegation position. His research interests include Computer Vision, Image Processing, and Pattern Recognition. He developed several systems for content-based image and video retrieval, focusing on interactive learning-based approaches. He is also interested in Machine Learning for Multimedia processing, Digital preservation, and Web archiving. Prof. Cord has published a hundred scientific publications and participated in several international projects (European FP6 and FP7, Singapore, Brazil) on these topics. He is a member of the IEEE.