GAZE LATENT SUPPORT VECTOR MACHINE FOR IMAGE CLASSIFICATION

Xin Wang *Nicolas Thome* *Matthieu Cord*

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

This paper deals with image categorization from weak supervision, *e.g.* global image labels. We propose to improve the region selection performed in latent variable models such as Latent Support Vector Machine (LSVM) by leveraging human eye movement features collected from an eye-tracker device. We introduce a new model, Gaze Latent Support Vector Machine (G-LSVM), whose region selection during training is biased toward regions with a large gaze density ratio. On this purpose, the training objective is enriched with a gaze loss, from which we derive a convex upper bound, leading to a *Concave-Convex Procedure* (CCCP) optimization scheme. Experiments show that G-LSVM significantly outperforms LSVM in both object detection and action recognition problems on PASCAL VOC 2012. We also show that our G-LSVM is even slightly better than a model trained from bounding box annotations, while gaze labels are much cheaper to collect.

Index Terms— weakly supervised learning, latent SVM, eye-tracking, gaze features, CCCP

1. INTRODUCTION

In the era of statistical machine learning, an overwhelming amount of images is available. For example, ImageNet [1] contains more than 10 million labeled images. The number of images also grows rapidly as multimedia communications and mobile devices become essential in our decades. Ever since 2013, 350 million new photos are posted on Facebook everyday [1]. We may never be able to label all images in a reasonable time by human labor.

The success of deep learning in computer vision largely improved the image classification accuracy on ImageNet [2]. Also, the deep models trained on large scale datasets were found to adapt well to other specific vision tasks [3, 4]. However, the problem is exacerbated when detailed annotations are used, *e.g.* bounding-boxes or pixel-wise labels, which are much onerous and expensive to collect than image-level annotations [5]. Additionally, using such accurate annotations during training can significantly boost categorization performances [6].

One option to get the best of both worlds is to develop weakly supervised learning (WSL) frameworks. Basically, WSL consists in designing accurate models able to predict detailed annotations, *e.g.* region localization, while being trained from coarse labels, *e.g.* global image labels. In machine learning, the Latent Support Vector Machine (LSVM) [7] introduces a theoretically sound formalism for WSL. Several attempts have been devoted to applying LSVM for weakly supervised object detection and scene recognition [8, 9, 10, 11, 12, 13]. One challenge with LSVM is due to the introduction of latent variables, which makes the resulting optimization problem non-convex. To improve training performances, different solutions have recently been explored to apply the curriculum learning idea, *i.e.* how to find easy samples to incrementally train the model [14, 10, 11]. When using sliding window approaches, the size of the latent space becomes huge. To overcome this issue, incremental exploration strategies have been proposed in [15, 10, 11]. Finally, recent works focus on enriching the prediction function, by using several (top) instance scores instead of using a single max [16], or by incorporating negative evidence [17, 18].

In this paper, we propose to incorporate gaze features extracted from an eye-tracker device, to improve the training of LSVM models. Gaze features are appealing since they can be generated by humans at almost zero-cost when performing a recognition task. Collecting gazes is more user-friendly and less time-consuming than collecting traditional annotations: it takes about 1 second to collect gazes for one image [19], comparing to 26s for drawing a bounding-box [20] and 15-60 min for labeling the segmentation mask for an image [21]. For different purposes, people design different collection protocols to acquire gazes [22, 19, 23, 24]. The collection protocols can be grouped into two main groups: task-driven and free-viewing. Task-driven means the annotators are given a specific semantic to look at, *e.g.* dog or a group of actions. Free-viewing means the annotators view the image freely. In this paper, we use two task-driven datasets [19, 23]. Recently, attempts have been devoted to incorporating gazes as weak supervision signals [25, 19, 26] for improving the performance of classification or segmentation systems. In [19], objects detectors are trained from gaze features instead of accurate bounding boxes, showing promising results.

This paper introduces G(aze)-LSVM, a new weakly supervised learning model for image classification. G-LSVM generalizes latent SVM (LSVM) by exploiting human gazes

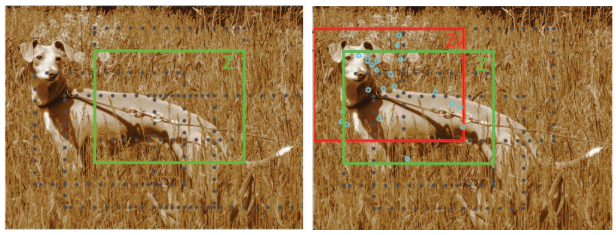
[1] http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9?IR=T

for localizing objects. Fig. 1 illustrates the rationale of our model, where latent variables correspond to all possible regions in the image. The goal of LSVM is to select semantically meaningful regions, *e.g.* those containing the target object class (region Z in Fig. 1a). To improve the quality of the region selection, G-LSVM supports regions with high density of gazes (region Z in Fig. 1b) with respect to the region with the highest density of gazes (region Z_i in Fig. 1b), by assuming that gaze features are related to regions relevant for the recognition task. Unlike [27, 25], G-LSVM only exploits gazes during training phase, and uses the pure visual information at test time without gazes. While [19, 28] focuses respectively on object detection and action recognition in the video, our targeted goal is to improve classification performances in the still images by using improved predicted latent regions.

In section 2, we formally define G-LSVM and its training procedure. Experiments conducted in section 3 show that G-LSVM significantly outperforms LSVM on both object dataset and action dataset of PASCAL VOC 2012.

2. GAZE-LSVM (G-LSVM) MODEL

We consider the problem of learning with weak supervision in a binary classification context.



(a) LSVM (b) G(aze)-LSVM

Fig. 1: Gazes bias the selection of latent regions for LSVM. The interpretation is in the section 1.

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

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

Note that our prediction function in Eq. (1) is the same as in LSVM [8], such that the label prediction at test time does not need gaze information.

2.1. G-LSVM Training

Our training scheme, however, penalizes the selection of latent regions based on gaze information. The general expression of G-LSVM training objective is as follows:

$$\mathcal{L}_G(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_c(\hat{y}_i, y_i) + \gamma \cdot \Delta_g(\hat{z}_i, z_i) \quad (2)$$

where y_i is the true label of image x_i, z_i is the region with the maximum number of gazes, $\hat{y}_i = \operatorname{sgn}(f_{\mathbf{w}}(x_i))$ is the label predicted by our model, $\hat{z}_i = \arg\max_{z \in Z(x_i)} \langle \mathbf{w}, \Phi(x_i, z) \rangle$ is the selected region, and $\frac{1}{2}\|\mathbf{w}\|^2$ is the standard max margin regularization term. For each training example, Eq. (2) includes a classification loss Δ_c, and a gaze loss Δ_g, with a trade-off parameter γ. A standard classification metric is the 0/1 loss, which is, however, difficult to optimize. As in SVM, we use the hinge loss as upper-bound, so that $\Delta_c(\hat{y}_i, y_i) = \max(0, 1 - y_i f_{\mathbf{w}}(x_i))$.

The novelty in our training scheme is the introduction of a gaze loss. Its preliminary definition is δ_g:

$$\delta_g(\hat{z}_i, z_i) = 1 - \frac{\mathrm{g}(x_i, \hat{z}_i)}{\mathrm{g}(x_i, z_i)} \quad (3)$$

where $\mathrm{g}(x_i, z)$ is the number of gazes in the region z for image x_i. Fig. 1b illustrates the proposed gaze loss, with blue circles representing gaze annotations, and z_i is shown in red. In this example the region z contains 18 gazes out of 20 for z_i, so that the gaze loss is 0.1, leading to a small penalization.

$\delta_g(\hat{z}_i, z_i)$ in Eq. (3) 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* [29]:

$$\Delta_g(\hat{z}_i, z_i) = \max_{z \in Z(x_i)} [\delta_g(z, z_i) + \langle \mathbf{w}, \Phi(x_i, z) \rangle] - \langle \mathbf{w}, \Phi(x_i, z_i) \rangle \quad (4)$$

Our training objective in Eq. (2) is thus biased by the gaze loss Δ_g, so that G-LSVM learns different $\mathbf{w}$ parameters compared to LSVM.

2.2. Optimization

To minimize our training objective function, we first show that Eq. (2) can be rewritten as a difference of convex functions, *i.e.* $u(\mathbf{w}) - v(\mathbf{w})$, where $v(\mathbf{w}) = C \sum_{i_p=1}^{n_p} f_{\mathbf{w}}(x_{i_p})$, $u(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i_p=1}^{n_p} [\gamma \Delta_g(\hat{z}_i, z_i) + \max(1, f_{\mathbf{w}}(x_{i_p}))]$ $+ C \sum_{i_n=1}^{n_n} \max(0, 1 + f_{\mathbf{w}}(x_{i_n}))$, where n_p (n_n) is the number of positive (negative) examples. Note that in the previous decomposition, the non-convex classification loss of every positive example is first decomposed into a difference of two convex functions: $\max(0, 1 - f_{\mathbf{w}}(x)) = \max(1, f_{\mathbf{w}}(x)) - f_{\mathbf{w}}(x)$.

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 [30]. 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 $\mathbf{v}_t = -\sum_{i_p=1}^{n_p} \Phi(x_i, \hat{z}_i)$. At line 4, the problem becomes convex, we can use any convex optimization tool for solving this problem, *e.g.* SGD.

Algorithm 1: Concave-Convex Procedure

Output: $\mathbf{w}^*$

Set $t = 0$, stopping criterion ϵ and initialize $\mathbf{w}$ by $\mathbf{w}_0$

repeat

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

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

  Set t = t+1,

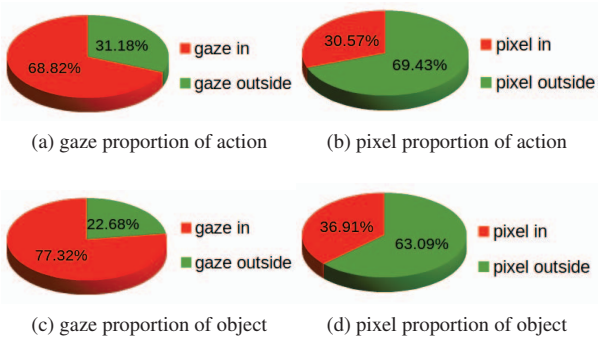
until $[u(\boldsymbol{w}_t) - v(\boldsymbol{w}_t)] - [u(\boldsymbol{w}_{t-1}) - v(\boldsymbol{w}_{t-1}))] < \epsilon$;

3. EXPERIMENTAL RESULTS

Two datasets, PASCAL VOC 2012 *action* and *object*, are used for evaluation. For both datasets, [23, 19] collected gaze annotations in task-driven manners.

3.1. Statistical consistency of gaze information

Before evaluating G-LSVM, we first provide a detailed analysis of the gaze data consistency. We compute statistics for the proportion of gazes falling into or outside the bounding boxes and compare it to the proportion of image pixels (Fig. 2). Statistically, for action dataset, 68.8% of the gazes fall into the ground-truth bounding-box, while the score of pixels is only 30.6%. Similarly, the scores of object dataset is 77.3% *vs* 36.9%. This preliminary study provides a quantitative validation that human gazes are highly related to object localization, and convey relevant features for classification.



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

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

Fig. 2: Proportions of gazes and pixel numbers in (outside) the ground-truth bounding boxes.

3.2. Experimental results

Setup: In this paper, latent regions correspond to square image regions extracted with a multi-scale sliding window strategy. Region sizes vary from 90% to 30% of the whole image area, with a stride of 10%. Each region is described by the state-of-the-art deep features extracted from the pre-trained *imagenet-vgg-m-2048* deep model[2], which are subsequently L2-normalized.

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

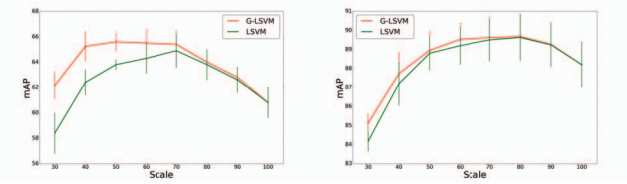
Performance comparison: We compare G-LSVM to the baseline LSVM, with $C = 10^4$ for both models, and $\gamma = 0.2$ for G-LSVM. G-LSVM and LSVM are trained independently for each scale. We perform scale combination using a simple object bank representation [31], leading to a 8-dimensional vector for each image. We also report performances of a SVM classifier trained on deep features computed on the whole image, denoted as wSVM.

The results are gathered in Table. 1, using 5 random folds on the `train+val` sets [32], and evaluating performances with the standard mAP metric. We show that G-LSVM outperforms LSVM by a margin of 2.1% for action (resp. 0.4% for object). Paired T-tests reveal that the improvement is statistically significant for a risk of less than 0.5% for action (resp. 2% for object). Both methods largely outperform wSVM, which clearly validate that training WSL models is able to capture local information.

	G-LSVM	LSVM	wSVM
action	70.5 ± 0.8	68.4 ± 1.0	60.8 ± 1.2
object	92.4 ± 1.0	92.0 ± 1.1	88.2 ± 1.2

Table 1: mAP(%) of combination multi-scale model.

Fig. 3 shows the performance evolution for LSVM and G-LSVM when varying the region scale s. We observe that the improvement of G-LSVM is more pronounced at small scales. This is expected: for large scales, all regions are informative, whereas at smaller scales, the model has to focus on relevant localized features. Note that s =100% corresponds to wSVM, for which the mAP for action and object is 60.8% and 88.2% . G-LSVM thus outperforms SVM at all scales of action dataset, as well for scales in $[50, 90]$ on object dataset.



(a) action dataset (b) object dataset

Fig. 3: mAP(%) at different scales.

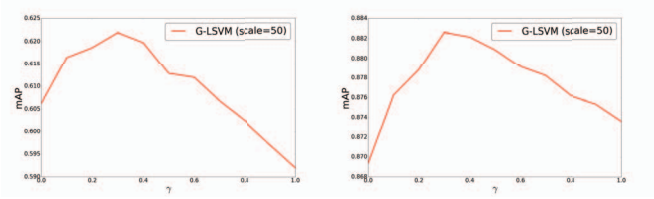
Table. 2 gives per-class performances at the smallest scale 30%. G-LSVM outperforms LSVM by a margin of 3.1% and 0.8% for respectively action dataset and object dataset. Paired T-tests show that G-LSVM is significant than LSVM for a risk less than 0.5% and 1% for action and object datasets. For action dataset, the performance gain of G-LSVM is especially large for the categories *phoning, reading, walking*. We note that these actions are usually associated with tiny objects, *i.e.* cellphone, book and small person (*e.g* Fig. 5b). For object dataset, G-LSVM performs well at *cow* and *motorbike* and improves over LSVM for most categories.

Further analysis: The impact of the parameter γ in Eq. (2) is shown in Fig. 4 for scale 50%. We can see that

Action Dataset	mAP	jump	phone	instru'	read	bike	horse	run	photo	comp'	walk
G-LSVM	**61.29**	**69.20**	**50.51**	**79.49**	**50.57**	78.86	83.88	53.62	**38.64**	72.08	**36.03**
LSVM-Standard	58.17	68.93	41.95	79.21	39.11	**79.26**	**84.20**	**55.11**	36.74	**73.69**	23.52
Object Dataset	mAP	aeroplane	cow	dog	cat	motor	boat	horse	sofa	din'table	bike
G-LSVM	**85.39**	**96.76**	**76.78**	**91.71**	**90.77**	**88.15**	**88.08**	82.82	**71.14**	**82.08**	**85.59**
LSVM	84.59	96.72	71.97	91.27	90.03	86.30	87.84	**84.05**	71.19	81.83	84.75

Table 2: AP(%) at scale 30%

performances of LSVM, corresponding to $\gamma = 0$, can be improved for most values in $\gamma \in]0, 1.0]$. It is worth noticing that the performances in Fig. 4 are shown on average for all classes. We can further substantially boost the performances by cross-validating γ. For example, on the action dataset, a class-wise cross validation ($\gamma \in [0, 1; 0.1]$) at scale 50% leads to nearly 1% improvement compared to $\gamma = 0.2$.

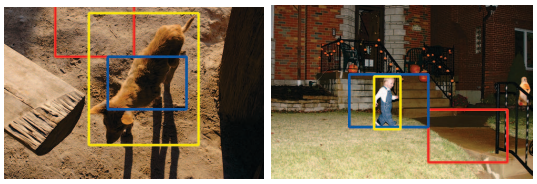


(a) action dataset (b) object dataset

Fig. 4: For scale = 50%, the effect of parameter γ.

We show in Fig. 5 the predicted regions for G-LSVM and LSVM. Results for training images are shown on the first row:

(a) dog (object) (b) walking (action)



(c) dog (object) (d) walking (action)

Fig. 5: Localization results. (a)(b): training results, (c)(d): test results. *red*: LSVM, *blue*: G-LSVM, *yellow*: ground-truth bounding-box. *cyan*: gazes.

we show that G-LSVM selects areas with more gaze features than LSVM. On the second row, we present results for test images, for which gaze features are not available. Interestingly, we can see that G-LSVM extracts regions which are more semantic than LSVM for the classification task.

We validate this idea by measuring the detection performances of G-LSVM *vs* LSVM by computing the Intersection over Union (IoU) metric between the predicted region and the ground-truth bounding boxes. The results in Table. 3 at every scale show that G-LSVM always outperforms LSVM.

action	30	40	50	60	70	80	90
G-LSVM	21.4	25.8	27.6	28.3	29.0	29.3	28.1
LSVM	14.5	20.4	24.3	26.7	27.9	28.9	28.0
object	30	40	50	60	70	80	90
G-LSVM	22.4	29.4	34.0	37.1	40.1	41.8	42.2
LSVM	20.1	27.1	32.6	36.4	39.2	41.5	42.0

Table 3: IoU (%) between predicted region and ground-truth bounding boxes.

Finally, we perform the last experiment using bounding box annotations during training, leading to a model denoted as G-LSVM*. We replace the gaze loss in Eq. (3) by a ground-truth loss computed as $1 - IoU(z, z_{gt})$, where z_{gt} is the ground-truth region in the dataset. The experiment reveal that G-LSVM is even slightly better than G-LSVM* ($\uparrow$ 0.4 % (0.2%) mAP for the action (object) dataset). This shows that gaze features contain as relevant information as bounding box annotations, while being much cheaper to collect.

4. CONCLUSION

In this paper, we introduce G-LSVM, a new latent variable model which leverages human gaze features during training. We derive a concave-convex upper bound of the non-convex problem and solve it by the CCCP. When gaze annotations are scarce, an appealing feature of G-LSVM is that the model only uses the gazes for training, whereas only visual information is used for prediction. Experimental results show that G-LSVM significantly outperforms LSVM on classification and localization tasks, and that the model achieves similar performances as a model trained with expensive bounding box annotations. In the future, we plan to research gazes as time-series for mining more information.

5. REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*. 2012.

[3] Xin Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *ICME workshop*, 2015.

[4] Marion Chevalier, Nicolas Thome, Matthieu Cord, Jérôme Fournier, Gilles Henaff, and Elodie Dusch, "LR-CNN for Fine-Grained Classification With Varying Resolution," in *ICIP*, 2015.

[5] Matthew Blaschko, Pawan Kumar, Ben Taskar, "Tutorial: Visual learning with weak supervision," in *CVPR*, 2013.

[6] Thibaut Durand, Nicolas Thome, Matthieu Cord, Sandra Eliza Fontes de Avila, "Image classification using object detectors," in *ICIP*, 2013.

[7] Stuart Andrews, Ioannis Tsochantaridis, Thomas Hofmann, "Support Vector Machines for Multiple-Instance Learning," in *NIPS*, 2002.

[8] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part Based Model," *PAMI*, 2010.

[9] Megha Pandey, Svetlana Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011.

[10] Olga Russakovsky, Yuanqing Lin, Kai Yu, and Li Fei-Fei, "Object-centric spatial pooling for image classification," in *ECCV*, 2012.

[11] Bilen, H. , Namboodiri, V.P. , Van Gool, L.J., "Object classification with latent window parameters," *IJCV*, 2014.

[12] M. Juneja and Jawahar C. V. Zisserman A. Vedaldi, A., "Blocks that shout: Distinctive parts for scene classification," in *CVPR*, 2013.

[13] Jian Sun and Jean Ponce, "learning discriminative part detectors for image classification and cosegmentation," in *ICCV*, 2013.

[14] P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *NIPS*, 2010.

[15] T. Durand, N. Thome, M. Cord, and D. Picard, "Incremental learning of latent structural svm for weakly supervised image classification," in *ICIP*, 2014.

[16] Weixin Li and Nuno Vasconcelos, "Multiple instance learning for soft bags via top instances," in *CVPR*, 2015.

[17] Thibaut Durand, Nicolas Thome, and Matthieu Cord, "MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking," in *ICCV*, 2015.

[18] Thibaut Durand, Nicolas Thome, and Matthieu Cord, "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks," in *CVPR*, 2016.

[19] Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, Vittorio Ferrari, "Training object class detectors from eye tracking data," in *ECCV*, 2014.

[20] Hao Su, Jia Deng, Li Fei-Fei, "Crowdsourcing Annotations for Visual Object Detection," in *AAAI Workshop*, 2012.

[21] Pushmeet Kohli, L'ubor Ladický, Philip H.S. Torr, "Robust Higher Order Potentials for Enforcing Label Consistency," *IJCV*, vol. 82, no. 3, pp. 302–324, 2009.

[22] Stephanie Lopez, Arnaud Revel, Diane Lingrand, Frederic Precioso, "One gaze is worth ten thousand (key-) words," in *ICIP*, 2015.

[23] Stefan Mathe, Cristian Sminchisescu, "Action from still image dataset and inverse optimal control to learn task specific visual scanpaths," in *NIPS*, 2013.

[24] S. Karthikeyan, Vignesh Jagadeesh, Renuka Shenoy, Miguel Eckstein, B.S. Manjunath, "From Where and How to What We See," in *ICCV*, 2013.

[25] Alireza Fathi, Yin Li, James M. Rehg, "Learning to Recognize Daily Actions using Gaze," in *ECCV*, 2012.

[26] Dimitris Samaras Gregory J. Zelinsky Gary Ge, Kiwon Yun, "Action Classification in Still Images Using Human Eye Movements," in *CVPRW*, 2015.

[27] Iaroslav Shcherbatyi, Andreas Bulling, Mario Fritz, "GazeDPM: Early Integration of Gaze Information in Deformable Part Models," *CoRR*, 2015.

[28] Leonid Sigal Greg Mori Nataliya Shapovalova, Michalis Raptis, "Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization," in *NIPS*, 2013.

[29] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu, "Cutting-plane training of structural svms," *Mach. Learn.*, 2009.

[30] A. L. Yuille, Anand Rangarajan, "The Concave-Convex Procedure (CCCP)," in *NIPS*, 2002.

[31] Li-Jia Li, Hao Su, Eric P. Xing, Li Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *NIPS*, 2010.

[32] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman, "The Pascal Visual Object Classes Challenge a Retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015.