MODELING HIERARCHICAL AND HETEROGENEOUS FEATURE REPRESENTATION WITH CONDITIONAL RANDOM FIELD FOR VISUAL OBJECT DETECTION

YaPing Zhu[a], *Steve Pan*[b] *and JianPing Chai*[a]

[a] Dept. of Communication Engineering, Communication University of China, Beijing, China, 100024
[b] School of Automation, Southeast University, Nanjing, China, 210096

ABSTRACT

We propose a novel flexible and hierarchical object representation using heterogeneous feature descriptors for detection of visual objects in real-world scenarios. Our representation is built on a Conditional Random Field (CRF) model that is able to aggregate local, semi-local and global features in one consistent framework. To improve the discriminative power of our model, we incorporate SVM classifiers into the CRF to learn discriminative unary classifiers for different object parts. Besides parameter learning of unary classifiers, a topology learning that captures the underlying geometrical structure of the target object class also boosts the performance of our model. Evaluation results on both simple UIUC single-scale car dataset and the challenging PASCAL VOC 2007 dataset verify that our model is flexible enough for a wide variety of object classes and robust to appearance variations caused by pose changes, articulation and partial occlusion.

Index Terms— Object detection, Conditional random field, Hierarchical representation.

1. INTRODUCTION

Visual object detection and categorical object classification is one of the primary yet sophisticated topics in computer vision and pattern recognition communities. Although much progress [1-11] has been made recently, it is still far from been well established for practical applications. Object detection is challenging because objects from the same class can vary significantly in appearance and shape. Variations arise not only from changes in illumination and viewpoint, but also due to articulation and intra-class variability in shape and other visual properties. In addition, object detection is a highly imbalanced classification task, which means that a typical natural image may contain many more negative background patterns than object patterns.

Recent developments have shown the effectiveness of using different feature types and hierarchical feature representations in the context of object detection and categorization. Zhang et al. [1] fused local texture features represented by PCA-SIFT and global shape context descriptors within a single multi-layer Adaboost model for object class recognition. Shotton et al. [2] introduced a method of fusing contour and texture information for categorical object detection. Combining advantages of shape and appearance features, we [3] proposed an effective approach for object categorization and detection, in which we built a generic shape codebook extracted from a set of pair adjacent segments and a class-specific appearance codebook selected from a pool of heterogeneous local descriptors. Since different kinds of feature descriptors provide complementary views of image data and increase the interpretability of object representation, all these works show a performance enhancement after combination of heterogeneous features. On the other hand, hierarchical feature representations [4-6] are a powerful paradigm for object detection. Usually, hierarchical representation presents informative views of object data at different scale levels, so it derives a flexible structure for capturing appearance, articulation and viewpoint changes. Consequently, extracting heterogeneous feature descriptors and organizing them in a hierarchical structure expect to improve the detection and classification performance.

Modeling and learning object representation is crucial in any object detection and classification system, be it human or computer. To date, several powerful machine learning and data mining models have been proposed, which makes the exploitation of huge feature data and complex spatial layout become practical. Boosting techniques encode objects' appearance codebook by selecting them from a large number of local features within the sample window. The relative positions of these local features represent the shape implicitly. Constellation and star model organize local parts and estimate their joint appearance-spatial distribution. These probabilistic part-based models have shown good performance on several benchmark datasets. However, the complexity of the combined estimation step restricts them to a relatively small number of parts and both models usually assume a fixed object parts so that the solution space is highly restricted. More recently, some complex models like Conditional Random Field (CRF) [4-8], and Latent SVM are proposed. As latent models allow to incorporate discriminative classifiers into a generative model, they facilitate modeling object categories with more deformable and interpretable structures. Liu et al. [7] established a CRF to learn and combine three kinds of features for still and dynamic salient object detection. Schnitzspan et al. [8] assembled an object based on a flexible ensemble of parts whose labels were treated as hidden nodes in a latent CRF. With the help of this latent CRF model, their method enables the automatic discovery of semantically meaningful object part representations. While Felzenszwalb et al. [9] tackled partially labeled training data using a similar model. They combined a margin-sensitive method for mining hard negative examples with a latent variable formulation called latent SVM which led to an iterative training algorithm that alternates between fixing latent values for positive examples and optimizing the latent SVM objective function. Regarding these

This work was supported by the National Natural Science Foundation of China under Grant 60805002, 90820009 and 61101165, and Foundation for Young Scholars of Southeast University under Grant 4008001015.

unlabelled parts as latent variables in object model, the resulting object detection system is formulated as mixtures of flexible part models.

Considering the strength of hierarchical and heterogeneous feature representation, as well as the flexible part-based object model, in this paper, we propose a novel object class detection approach that incorporates the aforementioned modules into a consistent CRF framework. In particular, we extract a set of complementary feature descriptors under different grid sizes to represent a generic object class in local, semi-local and global sense with a hierarchical and multi-feature manner. The arrangement of object parts is characterized and learned with a graphical-like CRF model, where each node (unary potential) associates with detected hierarchical features and edge (pairwise potential) defines the connection between two arbitrary nodes. With this flexible CRF-based object model, our approach is able to provide robustness to articulations and missing features caused by partial occlusion or pose changes. Evaluations on the two benchmark datasets demonstrate the effectiveness of our model.

2. MODEL FORMULATION

Given an image I, we formulate object detection as a binary classification problem, i.e. for I, a binary label $L\in\{1,-1\}$ indicate whether a particular class of objects contain in the image. Motivated by the recent success of CRF in categorization and verification tasks, we also model this problem using a pairwise graph structure-based CRF [4-8] that estimates the posterior probability $P(L|I)$. In particular, each node $n\in V$ in the graph $G(V,E)$ represents a binary part label $l_n\in\{1,-1\}$ encoding the presence of an object part from the target class (l_n =1) or a background patch (l_n =-1). The set of active edges E connecting the nodes defines the interaction between pairwise nodes and edge structure of the underlying CRF. To facilitate the classification of object parts from image features, our model integrates discriminative SVM classifiers into generative CRF framework. In addition, a topology learning that explores the inherent geometrical relationship and parts structure for object class also contributes to the high performance of our system.

2.1. CRF-based object model

Since in object detection we are interested in the probability of presence or absence of particular objects, the posterior distribution can be modeled as

$$P(L\mid I;\boldsymbol{\theta})=\frac{1}{Z(\boldsymbol{\theta})}\prod_{n\in V}u_n(l_n,I;\boldsymbol{\theta})\prod_{(m,n)\in E}v_{mn}(l_m,l_n,I;\boldsymbol{\theta}) \qquad (1)$$

where $\boldsymbol{\theta}=\{\boldsymbol{\alpha},\ \boldsymbol{w},\ \boldsymbol{e}\}$ parameterizes the CRF model, $Z(\boldsymbol{\theta})$ is a normalization function, u_n and v_{mn} denote the unary and pairwise (edge) potentials. Specifically, $\boldsymbol{\alpha}$ and $\boldsymbol{w}$ are parameters of unary potentials, whereas, $\boldsymbol{e}$ is parameter of pairwise potentials.

In our model, unary potentials collect local, semi-local and global evidence. Evidence of each u_n corresponds to a specific feature descriptor f_n (I). As shown in Fig.1, we describe an image with a hierarchical structure, where unary potentials at bottom level extract local features and potentials at intermediate levels attain semi-local features in a larger region until a global representation of the object is aggregated at the top level. Similar to [8], we define the unary potential of node n using a softmax function based on a weighted combination of the response of a group of M hybrid feature classifiers $\boldsymbol{F}(\boldsymbol{\alpha}_n f_n(I))=[F(\boldsymbol{\alpha}_{n,1},f_n(I)),\ \ldots,$

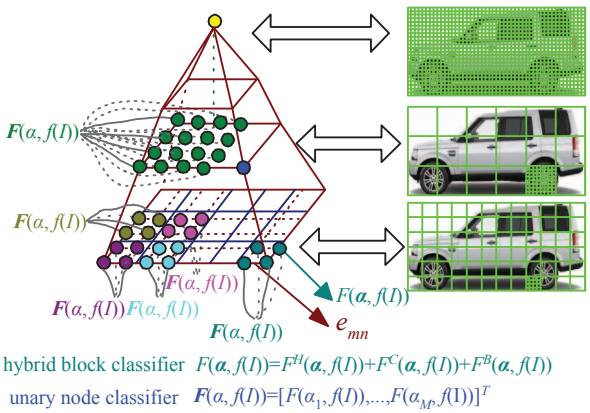

Fig.1. Framework of our CRF-based hierarchical object model

$F(\boldsymbol{\alpha}_{n,M},\ f_n(I))]^T$, where $\boldsymbol{w_n}$ are the weight vector, and $\boldsymbol{\alpha_n}$ are the parameters of classifiers.

$$u_n(l_n,I;\boldsymbol{\alpha},\boldsymbol{w})=\frac{\exp(l_n\cdot\boldsymbol{w}_n^T\boldsymbol{F}(\boldsymbol{\alpha}_n,f_n(I)))}{\sum_{l\in\{-1,1\}}\exp(l\cdot\boldsymbol{w}_n^T\boldsymbol{F}(\boldsymbol{\alpha}_n,f_n(I)))} \qquad (2)$$

Pairwise potentials extract the topology of objects as they model the connection between node pair (m, n). Considering the interaction of image features $f_m(I)$ and $f_n(I)$, we define the pairwise potentials based on a similar softmax function on concatenated unary features

$$v_{mn}(l_m,l_n,I;\boldsymbol{e})=\frac{\exp\left(\left[f_m(I),f_n(I)\right]^T\boldsymbol{e}^{mn}_{l_m l_n}\right)}{\sum_{\{a,b\}\in\{-1,1\}}\exp\left(\left[f_m(I),f_n(I)\right]^T\boldsymbol{e}^{mn}_{ab}\right)} \qquad (3)$$

For each possible edge (m,n) and each combination of labels $(a,b)\in\{-1,1\}$, we use a particular classification vector $\boldsymbol{e}^{mn}_{ab}$, which allow us to model spatial dependencies and layouts of different feature types.

It is worth noting that by independently classifying nodes based on corresponding unary potentials and jointly encoding dependencies among neighboring nodes according to pairwise potentials, our model is capable of providing a discriminative multiscale and complementary view on objects and simultaneously capturing the underlying topology of objects. This is the difference between our model and previous works.

2.2. Hierarchical and heterogeneous feature representation

Three types of heterogeneous feature descriptors like Histograms of Oriented Gradients (HOG) [10], Center-Symmetric Local Binary Pattern (CS-LBP) [11], and Binary Coherent Edge descriptors (BiCE) [12] are used to describe the contents in an image cell. In particular, HOG and CS-LBP are gradient orientations-based and texture-based descriptors, respectively, whereas BiCE is edge-based descriptor which encodes the presence or absence of edges using a binary value for a range of possible edge positions and orientations. These heterogeneous descriptors represent an image patch from different aspects, thus providing a complementary view of objects.

For each kind of descriptor, to compute a hierarchical feature representation, at the bottom level, we segment the whole image into a dense grid of non-overlapping $i\times i$ pixel cells and concatenate neighboring $j\times j$ cells to one block. For higher levels, we successively double the size of blocks in horizontal and vertical

directions until on the highest level. By calculating the corresponding descriptors at each block and concatenating several adjacent blocks to form feature descriptors at different levels, we attain a hierarchical representation that consists of the local, semi-local and global features for the target object. Following the construction in [10-12], for HOG and CS-LBP, (i, j) is set to (8,2) and (4,1), respectively, whereas, for BiCE, i is set to 8 and j is set to 3. Finally, to remove background noise and statistical redundancy in the feature space, we apply PCA on each kind of feature descriptor which result in the final 11-dimensional PCA-HOG, 60-dimensional PCA-CS-LBP (we use CS-LBP$_{2,8,0.01}$ in this paper) and 256-dimensional PCA- BiCE descriptor per block.

We train a discriminative χ^2 kernel SVM classifier for each kind of descriptor per block, and aggregate three descriptor-specific SVM classifiers($F^H(\boldsymbol{\alpha}, f(I))$, $F^C(\boldsymbol{\alpha}, f(I))$ and $F^B(\boldsymbol{\alpha}, f(I))$ denote PCA-HOG, PCA-CS-LBP and PCA- BiCE SVM classifier, respectively) into one hybrid block SVM classifier $F(\boldsymbol{\alpha}, f(I))$. To simplify the complexity of our representation, we merge several neighboring blocks into a unary node according to the number of scale levels in the hierarchical model. For each unary node n, the final unary SVM classifier $\boldsymbol{F}(\boldsymbol{\alpha}_n, f_n(I))$ is formulated as the concatenation of the hybrid block SVM classifiers in that node, i.e. $\boldsymbol{F}(\boldsymbol{\alpha}_n, f_n(I)) = [F(\boldsymbol{\alpha}_{n,1}, f_n(I)),\ldots,F(\boldsymbol{\alpha}_{n,M}, f_n(I))]^T$.

3. MODEL LEARNING

As discussed in sec.2.1, the parameter set $\boldsymbol{\theta}=\{\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{e}\}$ in our model consists of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{w}$ of unary potentials, as well as parameters $\boldsymbol{e}$ of pairwise potentials. Given K training images $\boldsymbol{I}=\{I^1,\ldots,I^K\}$ annotated with bounding boxes and their labels $\boldsymbol{L}=\{L^1,\ldots,L^K\}$, the goal of model learning is to optimize the parameter set $\boldsymbol{\theta}$ and decide a suitable graph structure representing the underlying geometrical topology of the object class.

3.1. Parameter learning

Like maximum likelihood training used in most CRF model [4-7], we optimize parameters such that a conditional log-likelihood $\ell(\theta) = \sum_{k=1}^{K} \log P(L^k \mid I^k;\boldsymbol{\theta}) + P(\boldsymbol{\theta})$ is maximized, where $P(\boldsymbol{\theta}) = P(\boldsymbol{w})\cdot P(\boldsymbol{e})$ is a regularizer that avoids overfitting.

The training of parameter $\boldsymbol{\alpha}$ for SVM classifiers is facilitated by the primal SVM training proposed in [13] which showed competitive results compared with quadratic programming in the dual form. We follow this idea and embed primal SVM training in the CRF model. Other parameters can be optimized via gradient ascent. In particular, assuming $P(\boldsymbol{w})\sim\mathcal{N}(0,1)$ and $P(\boldsymbol{e})\sim\exp(-\|\boldsymbol{e}\|)$, differentiating $\ell(\theta)$ w.r.t. $\boldsymbol{w}$ gives

$$\frac{\partial \ell(\theta)}{\partial \boldsymbol{w}_n} = \sum_{k=1}^{K} E_{\tilde{P}(\boldsymbol{L}|I^k;\theta)}\left[\boldsymbol{F}(\boldsymbol{\alpha}_n, f_n(I^k))\cdot l_n \cdot u_n(l_n, I^k)\right] - E_{P(\boldsymbol{L}|\boldsymbol{I})}\left[\boldsymbol{F}(\boldsymbol{\alpha}_n, f_n(I))\cdot l_n \cdot u_n(l_n, I)\right] - \boldsymbol{w}_n \quad (4)$$

where $E_{\tilde{P}(\boldsymbol{L}|I)}[\cdot]$ is expectation under empirical distribution of training data, and $E_{P(\boldsymbol{L}|\boldsymbol{I})}[\cdot]$ is the expectation under the posterior probability of our model. Similarly, taking partial derivatives of $\ell(\theta)$ w.r.t. $\boldsymbol{e}$ gives

$$\frac{\partial \ell(\theta)}{\partial \boldsymbol{e}^{mn}_{l_m l_n}} = \sum_{k=1}^{K} E_{\tilde{P}(\boldsymbol{L}|I^k;\theta)}\left[[f_m(I^k), f_n(I^k)]^T \cdot v_{mn}(l_m, l_n, I^k)\right] - E_{P(\boldsymbol{L}|\boldsymbol{I})}\left[[f_m(I), f_n(I)]^T \cdot v_{mn}(l_m, l_n, I)\right] - \mathrm{sgn}(\boldsymbol{e}^{mn}_{l_m l_n}) \quad (5)$$

To evaluate Eq.(4) and (5), it is required to compute the marginal distributions $P(l_n|I)$ and $P(l_m, l_n|I)$. Since these marginals do not have closed-form solution, we approximate them using loopy belief propagation.

3.2. Topology learning

Previous results [8, 14-15] show that learning a suitable geometrical topology greatly improves the classifier's performance for deformable and articulated objects. To this end, we also proposed to boost the flexibility of our model by learning parts structure and object topology. The intuition behind is to find a set of edges $E=\{(m^*,n^*)\}$ with the largest log-likelihood gradient $\partial \log P(l=1|I;\boldsymbol{\theta})/\partial \boldsymbol{e}^{m^*n^*}$, since theses edges best improve the discrepancy between object (l=1) and background (l=-1). Concretely, we iteratively search edges (m^*,n^*) that maximize the log-likelihood ratio as shown in Eq.(6), and add them to an initially empty set E from all possible edges $(m,n)\in V\times V-E$

$$(m^*,n^*) = \arg\max_{(m,n)\in V\times V-E} \left\| \frac{\partial \ell(\theta)}{\partial \boldsymbol{e}^{mn}_{11}} - \frac{\partial \ell(\theta)}{\partial \boldsymbol{e}^{mn}_{-1-1}} \right\| \quad (6)$$

After each iteration, we also remove those edges with absolute weight below a threshold τ_1 or with an absolute gradient norm below a threshold τ_2.

4. EVALUATION RESULTS

For the detection task, we evaluated our model on UIUC multi-scale car dataset [14]. For this dataset, we trained our model on 300 bounding boxes containing positive examples and 250 negative examples, randomly selected from training set. To evaluate the detection performance, we follow the detection procedure and evaluation protocol in [14] (see Eq.(2) in [14]). We compare equal error rate (EER) with several other works [3, 14-15]. In particular, 2AS-SVM [3] is our previous model which combines generic shape codebook with class-specific appearance codebook, and others are state-of-the-art methods. Fig. 2 plots the precision-recall curves for all methods. As shown that our CRF-based model achieves competitive result with EER = 95.54% compared to other well performing methods. Although, thanks to the combination of shape and appearance features, the old 2AS-SVM model by itself performs already quite well on this dataset (with EER = 91.31%), our new model still improves the average precision performance by 5.87%.

To further show the robustness of our model to appearance and shape deformation due to scale and pose changes as well as partial occlusion and articulation, we also provide results on more difficult PASCAL VOC 2007 challenge dataset [16] which has 20 object classes. We follow the PASCAL challenge protocol and report the average precision (AP). We compared our complete model (hierar + hetereo) with hierarchical and heterogeneous feature representation to the model relying only on hierarchical HOG descriptors (hHOG), or using only single-scale (the finest scale) heterogeneous features (hetereo), and other benchmark works. We summarize AP results on all 20 classes in PASCAL VOC 2007 dataset in Table 1. Compared with other works, our complete model achieves the best result on 14 of 20 listed classes. The average AP performance of complete model outperforms that of Felzenszwalb et al. [9] and Schnitzspan et al. [8] algorithm by 5.89% and 7.26%, respectively. Moreover, using hierarchical and heterogeneous feature descriptors improves the AP performance of using only single-scale heterogeneous descriptors and using only

Table 1. Comparison of AP (%) on 20 classes in PASCAL VOC 2007 dataset

Class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
hierar+hetereo	**35.4**	**57.2**	**9.7**	15.3	27.4	**45.7**	49.0	**24.8**	15.9	**20.5**
hHOG	33.5	57.0	7.6	14.8	26.8	45.1	47.5	23.7	13.0	20.2
hetereo	32.3	56.6	4.4	13.1	26.6	43.9	47.0	23.1	10.6	20.8
Felzenszwalb et al.[9]	32.8	56.8	2.5	**16.8**	**28.5**	39.7	**51.6**	21.3	**17.9**	18.5
Schnitzspan et al. [8]	31.9	57.0	9.1	15.2	26.0	42.7	49.3	14.5	15.2	18.5
	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
hierar+hetereo	24.3	**12.6**	**49.5**	**43.0**	36.2	**14.9**	**21.6**	**26.2**	**41.4**	**45.5**
hHOG	23.5	11.8	47.9	41.3	35.8	13.7	19.9	25.9	39.3	42.4
hetereo	23.2	11.6	47.1	40.9	35.6	13.2	20.4	24.5	39.5	43.6
Felzenszwalb et al.[9]	**25.9**	8.8	49.2	41.2	**36.8**	14.6	16.2	24.4	39.2	39.1
Schnitzspan et al. [8]	24.2	11.8	49.1	41.9	35.7	14.5	18.9	23.3	34.3	41.3

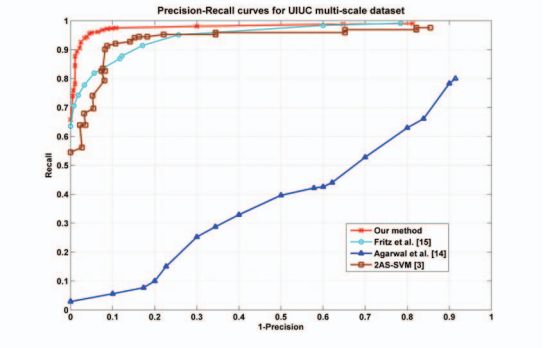

Fig.2. Precision-Recall curves for UIUC multi-scale dataset

Fig.3. Some detection examples on PASCAL VOC 2007 dataset

hierarchical HOG by 6.41% and 4.30%, respectively. Thereby, simultaneously including multiscale and complementary feature descriptors helps our framework to model a variety of complex object classes and also improves the detection performance. Fig. 3 shows some examples resulted from our detector.

5. CONCLUSION

In this paper, we propose a novel flexible part-based model to detect and classify visual object class in real-world scenarios. To conclude, our contribution includes 1). represent object classes using hierarchical and heterogonous features and model object representation with a consistent CRF framework. 2) in addition to the discriminative parameter learning, the proposed topology learning that captures implicit geometrical structure of object classes further helps to improve the performance of our model. Evaluation results on simple UIUC single-scale car dataset and the challenging PASCAL VOC 2007 dataset both verify the advantage of our new model compared with other state-of-the-art works.

6. REFERENCES

[1]. W.Zhang, B.Yu, G.J.Zelinsky and D.Samaras. Object class recognition using multiple layer boosting with heterogeneous features. *CVPR*, (2):323-330, 2005.

[2]. J. Shotton, A. Blake and R. Cipolla. Efficiently combining contour and texture cues for object recognition. *BMVC*, 2008.

[3]. H.Pan, Y.P.Zhu, L.Z.Xia and T.Q.Nguyen. Combining generic and class-specific codebooks for object categorization and detection. *ICASSP*, 2264-2267, 2011.

[4]. L.Ladicky, C.Russell, P.Kohli, P.Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. *ICCV*, 739-746, 2009.

[5]. P. Schnitzspan, M. Fritz, S. Roth and B. Schiele. Discriminative Structure Learning of Hierarchical Representations for Object Detection. *CVPR*, 2238-2245, 2009.

[6]. P.Schnitzspan, M.Fritz and B. Schiele. Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features, *ECCV*, (2):527-540, 2008.

[7]. T.Liu, Z.Yuan, J.Sun, J.Wang, N.Zheng, X.Tang and H.Shum. Learning to Detect A Salient Object. *IEEE Trans. PAMI*, 33(2):353-367, 2011.

[8]. P. Schnitzspan, S. Roth and B. Schiele. Automatic Discovery of Meaningful Object Parts with Latent CRFs. *CVPR*, 121-128, 2010.

[9]. P.Felzenszwalb, R.Girshick, D.McAllester and D.Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE PAMI*, 32(9):1627-1645, 2010.

[10]. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. in *Proc. CVPR*, (1):886-893, 2005.

[11]. M. Heikkila, M. Pietikainen and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425-436, 2009.

[12]. C. L. Zitnick. Binary coherent edge descriptors. *ECCV*, (2):170-182, 2010.

[13]. O.Chapelle. Training a Support Vector Machine in the Primal. *Neural Computation*, 19(5):1155-1178, 2007.

[14]. S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in image via a sparse, part-based representation. *IEEE Trans. PAMI*, 26(11):1475-1490, 2004.

[15]. M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. *ICCV,* (2):1363-1370, 2005.

[16]. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL VOC Challenge 2007.