Modeling Image Context using Object Centered Grid

Sobhan Naderi Parizi*, Ivan Laptev[†], Alireza Tavakoli Targhi*

**Computer Vision and Active Perception Laboratory*

Royal Institute of Technology (KTH)

SE-100 44, Stockholm, Sweden

{sobhannp,att}@kth.se

[†]INRIA / École Normale Supérieure

Paris, France

ivan.laptev@inria.fr

Abstract—Context plays a valuable role in any image understanding task confirmed by numerous studies which have shown the importance of contextual information in computer vision tasks, like object detection, scene classification and image retrieval. Studies of human perception on the tasks of scene classification and visual search have shown that human visual system makes extensive use of contextual information as post-processing in order to index objects. Several recent computer vision approaches use contextual information to improve object recognition performance. They mainly use global information of the whole image by dividing the image into several pre-defined subregions, so called fixed grid. In this paper we propose an alternative approach to retrieval of contextual information, by customizing the location of the grid based on salient objects in the image. We claim this approach to result in more informative contextual features compared to the fixed-grid based strategy. To compare our results with the most relevant and recent papers, we use PASCAL 2007 data set. Our experimental results show an improvement in terms of Mean Average Precision.

Keywords—Context; Scene; Histogram; Bag-of-Features;

I. INTRODUCTION

Contextual representation of natural images has recently turned into an interesting field of study in computer vision. Torralba uses contextual information to predict location and size of different classes of objects in an image [1]. Once this prior information is captured, one can run object detectors only on the promising locations of the image and make the detection process faster by saving a lot of computation [2]. In [3] context is used in conjunction with intrinsic object detectors to increase performance of the detector and enhance position and size of the detected bounding boxes. Marszałek et al. [4] have shown that many of human actions in movies can be classified better in the context of visual scenes. Another field of study where context plays the most critical role is classification of scene categories. In [5] [6] it has been shown that different images of an individual scene category share a lot of information in their global representation.

People take different approaches to exploit the contextual information of images. Sometimes, image background at the

local surroundings of the object is used as the context of the object [7]. Similarly, it has been shown that surroundings of object bounding box contain informative support for classification of animals [8]. Oliva et al. [9] model the context with a holistic representation over the whole image. Some others [10] [11] model the context by using the relation between objects and estimate the probability of co-occurrence of different objects and their correspondence constraints.

Several recent computer vision methods make benefit of histograms [12]. The simplicity of the histogram-based methods associated with their effectiveness in a wide variety of problems has made them very popular and widely used. One of the most general implementations of histogram-based methods is referred to as Bag of Features (BoF) [13].

Despite the enormous use of histogram-based methods, they frequently suffer from the loss of spatial information. Several papers address this problem. Lazebnik et al. introduces spatial pyramid idea where the image is subdivided into smaller regions several times and BoF for each region is calculated [6]. Marszałek et al. [14] create a fixed-grid over the whole image as shown in Figure 1. For each grid cell, one BoF is calculated independently. Then these BoFs are concatenated together to form the final feature vector. This method does not store the spatial information explicitly, however, an independent BoF models distribution of each spatial region of the image preserving the coarse spatial information. Some others explicitly add the x, y coordinates of each cell to the end of feature vector in order to retain the spatial information instead [5].

The limitation with all of these mentioned grid-based methods is that they define a *fixed* grid over the image. Keeping the grid fixed according to the image frame causes the method to be variant to image translation. However, the fixed-grid idea can model typical scene layouts in a fairly efficient way. As we will show later in Section IV, it performs much better than having no grid at all.

We suggest a new configuration for placing the grid on the image so that the spatial information is preserved and

besides the extracted feature is robust to image translation. The idea is to displace the grid according to position of the objects in the image. As an example, consider the category of images containing a scene of sea in the scene classification problem. If we try to adjust the grid such that the center cell always lies over the sea, we can most likely expect to have sky at the top cells and beach (or land generally speaking) at the bottom cells. More generally, we propose to modify the size and the position of the fixed-grid such that the center cell of the grid fits to the bounding box of one or more objects which are present in the image. Thus, we call the proposed method Object Centered Grid (OCG).

Our main contribution in this paper is, therefore, to show that the feature vector computed from OCG classifies the scenes better than fixed grid method. We claim that OCG will form the scene features in a much more coherent and informative way than the fixed-grid strategy. Our final goal here is not to beat the state-of-the-art method but to measure the maximum extent of gain in terms of average precision that one can achieve by using OCG compared to the fixed-grid counterpart.

The rest of the paper is organized as follows: in the next section we briefly address the concept of context in computer vision as well as the basics of our OCG idea. In Section III we explain the OCG method in more details. Section IV contains evaluations of the OCG method. We conclude the paper in Section V.

II. CONTEXT AND SCENE

Torrallba specifies that context is beneficial in object recognition tasks, in particular, when the intrinsic object features are not clear enough [1]. In other words, for images where the target object can be detected accurately enough by means of a local detector, it is preferable to rely on the result of the detector.

Contextual information comes into play when the object is locally hard to be detected. For example, street context is a strong cue to predict existence of a car within the image. Even if the car is highly occluded or, for whatever reason, the local information fails to single the car out, contextual information can still suggest the existence of a car.

A. Fixed-Grid vs. OCG

The main advantage of using fixed-grid scenario in building histograms is to incorporate spatial information into BoF framework. In the fixed-grid framework, the image is divided into different spatial segments by putting a grid on it (Figure 1). Each cell models the corresponding part of the image independently of other cells. BoF histograms calculated for grid cells are then concatenated into a single feature vector describing the image. Spatial information is implicitly stored in the fixed-grid representation given that different spatial regions are encoded by different parts of the feature vector.

Figure 1. One idea for incorporating spatial information in the BoF framework is to make a fixed-grid on the top of the image. An independent BoF is computed for each grid cell and then all of the BoFs are concatenated together in a fixed order. According to the image, the cells at the upper row are expected to cover sky and roof of the buildings; the middle cells are expected to cover a mixture of road, sidewalk and buildings; and the bottom cells are expected to mainly cover the road.

If considering example in Figure 1, what if the image would be translated such that it would not contain much of the sky (Figure 2a)? On the contrary, what if it would be translated in a way that it would not contain much of the road (Figure 2c)?

In the samples shown in Figure 2, it is no longer true that the BoFs attained from the upper cells represent the sky or buildings' roofs (Figure 2a). Similarly, in Figure 2c, the middle-row cells represent sky and the buildings' roof while we ideally expect to observe those regions in the topmost cells (cp. Figure 1). The same contradiction holds for the bottom cells of the Figure 2c.

In general, the fixed-grid idea works best when the images are fairly accurately aligned. In other words, if the goal is to classify different scenes, image alignment would mean to have the sea, buildings, dining table in the middle of the image for the sea scene class, street scene class, and kitchen scene class respectively.

The OCG idea suggests modifying the grid according to the objects within the image instead of expecting the images to be already aligned. Figures 2b, 2d show the OCG configuration while assuming the central object to be the car. Obviously, if the image is translated to any side, the grid cells will still represent almost the same region of the image in term of visual appearance. It means that if we use OCG approach, regardless of the image content, it is always expected to see road within the bottom cells, sidewalk at the middle cells and buildings and sky at the topmost cells.

Of course, one cannot expect the cells of an object-centered grid to always have a clean and unique content. However, we claim that OCG produces more coherent feature representation of image context than a fixed-grid representation since it is localized according to a real object. For example, it is not likely that the bottom cells of an OCG

Figure 2. This figure compares configuration of the grid in OCG and fixed-grid strategies. In (a) and (c), fixed-grid is used. (b) and (d) show the same images with OCG. Comparison of the OCG and fixed-grid in these two samples shows that OCG is much more invariant to translation and image content than fixed-grid.

contain buildings when we have a car in the center cell of the OCG.

III. OUR APPROACH

The main target of our work is to measure the amount of information that one could gain by incorporating the spatial information into the BoF framework through OCG strategy. Our results in Section IV show that the information gain of image context is considerably affected by the way that the context is modeled. Finally, we show that OCG forms the features in a coherent way such that contextual information of the image holds a fairly stable correspondence with respect to the central object. Unlike OCG, formation of the contextual features in the fixed-grid representation is more of a random nature.

We evaluate contextual gain on image classification problems where we aim to classify images into a set of object category classes (e.g. car, dining table, boat). For each pair of a test image and an object category we predict if the image contains any instance of that object regardless of its position in the image.

Image classification problem is dependent on the holistic scene representation of the image. For example, a car is most likely seen in a street scene; similarly, a dining table is most likely seen in a kitchen scene. It has been already shown that the fixed-grid idea works fairly promising in classifying scene categorization problems [6]. Therefore, the image classification problem suites well for demonstrating the performance gain that we can get with OCG method comparing to the fixed-grid alternative. Note, however, that the OCG method can be used for the scene classification task as well if we model scenes by OCG based descriptors centered on relevant objects such as chairs and tables for office scenes, trees and benches for park scenes, tents and people for outdoor market scenes and so on.

In this section we will first explain localization of OCGs according to the image frame and object bounding boxes followed by the feature representation. We describe our classification approach in Section III-C.

A. Localizing The OCG

As we explained earlier, the idea behind OCG is to partition the image with respect to position and size of some object of attention. We want the center cell of the grid to be placed over the object so that the other cells model the object surroundings.

We form the OCG in such a way that the center cell will exactly fit the bounding box of that object. For the evaluations in this paper, we use ground truth annotation of objects that are present in the image to form our OCGs. In this way we can measure the maximum amount of information gain we can expect from the OCG method. For images with more than one object of the same class we extract an individual OCG feature based on each of the objects. We classify an image by maximizing the classification scores of its OCG descriptors.

For negative samples, we do not have any objects of the corresponding class with annotated bounding boxes. One simple solution to construct OCG representation for negative samples is to randomly generate bounding boxes from a uniform distribution and localize our OCG with respect to them. According to [1], however, probability of having an object in a specific location of an image is not equally distributed over the image. We estimate a probability distribution for the position of different objects in the image as well as the expected size of the object bounding box independently of the image content. Our model for bounding boxes of objects is a multivariate gaussian distribution. The variables include x , y coordinates of the upper left corners of the bounding box as well as its width and height. We train parameters of this model from object bounding boxes in the training set.

Figure 3 illustrates the gaussian models that we obtained for spatial distribution of bounding boxes for object classes *bottle* and *car*. The images of Figure 3 are obtained in the following way: Firstly all of the pixels of the image are set to zero. Then 100 bounding boxes are sampled from the gaussian model that has been trained for the spatial distribution of the object. Finally, for each sampled bounding box, the intensity value of all of the pixels inside the bounding box is increased by one.

Figure 3. We model spatial distribution of different objects by multivariate Gaussian distributions. (a) represents distribution of location, size and aspect ratio of 100 random bounding boxes sampled from spatial distribution of *bottle*. Similarly, (b) represents bounding boxes sampled from spatial distribution of *car*.

B. Feature Extraction

It is believed that contextual information can be represented by the following types of features [1]: Distribution of structuring elements, color distribution and semantic relations.

Distribution of structuring elements: The structure of surrounding objects can help in classification of context into different scene classes. *Gist* is a type of feature that is useful for representing contextual information of an image and it has been used for scene classification problems [1] [2] [15]. *Gist*, in fact, is the response of gabor filters in several orientations and frequencies.

SIFT-like features have also been successfully used for modeling contextual information [6] [16]. In particular [5] has shown that Color-SIFT achieves relatively best performance for classification of scene categories. Moreover, [17] and [5] have shown that dense features perform better than sparse sampling for scene classification problems.

In order to model the structural distribution of context we create a BoF histogram over a visual vocabulary of HOG features [18] which are similar to SIFT [19]. The visual vocabulary is created by clustering the HOG features using K-Means algorithm. In order to accumulate BoF histogram, HOG features are extracted from overlapping image patches obtained by a regular sampling over positions and scales of the image. Each patch is labeled according to the index of the nearest match in the visual vocabulary and the BoF histogram of patch labels is accumulated. We refer to this histogram as *shape-BoF*.

Color distribution: Distribution of color features can also be informative in scene classification. For example, the dominant color in a forest scene is expected to be green while a beach scene is more bluish.

We model the color distribution in two levels. In the first layer each pixel of the image is labeled based on its RGB values. We use K-Means to build the visual vocabulary of

RGB values for the first layer. We call this vocabulary *CV1* and label all pixels in the image according to *CV1*. Next, similar to the shape-BoF representation, we consider regularly sampled overlapping image patches and compute color-patch descriptors as a histogram of *CV1* pixel labels. We use K-Means once more to create a new visual vocabulary for our color-patch descriptors. We call this new vocabulary *CV2* and use it to assign color label for each considered image patch.

The color distribution is, finally, modeled by accumulating a histogram over the computed patch labels according to *CV2*. We refer to this histogram as *color-BoF*.

Semantic relations: Co-occurrence of the set of objects that are present in an image can be considered as yet another source of information that can help us in scene classification. In [11], this type of information is used to re-rank the segment labels within an image. Moreover, one could also use spatial relations of the objects such as sky is on the top of the sea, road is underneath the car, etc. Heitz et al. automatically train the optimal set of spatial relations between objects [7]. Nevertheless, [20] claims that mutual location of the objects as a high-level representation cannot do much more than the low-level features in recognition of context.

As mentioned before, we use fixed-grid idea as our baseline. Fixed-grid incorporates the spatial co-occurrence relations in a simple way where the co-occurrence relations can be interpreted as follows: if the sky is at the top row grid-cells, buildings are at the middle row grid-cells, and the road is at the bottom row grid-cells, then the image most likely contains a street scene.

Our OCG method, exploits the semantic relationship between objects similar to the fixed-grid idea, however, due to the object-centered localization, we expect OCG to model the spatial information and semantic relations more coherently compared to the fixed-grid approach.

C. Classification

As illustrated in Figures 1, 2, we partition the image into 9 cells; both in OCG and fixed-grid strategy. We calculate one shape-BoF and one color-BoF for each cell and then concatenate them into a single feature vector.

It has been shown that in scene classification SVM is the superior classifier compared to KNN [5]. Therefore we use SVM for classification of our feature vectors. We train a SVM for each category of images using one-vs-all approach using RBF kernel. Our experiments verify the fact that χ^2 distance performs better than Euclidian distance in BoF framework [16] [21] [22]. In our scene classification experiments χ^2 increased the average precision scores from 5 to 15 percent depending on the object class when compared to Euclidian distance.

IV. EXPERIMENTAL RESULTS

As mentioned earlier, we approach the image classification problem from scene classification point of view. The actual problem to be solved is to answer the question whether there is an instance of a specific object class (say car) in an image or not. The way we answer this question is to make decision based on the contextual information of the image with respect to specific objects.

A. Dataset

We perform experiments on PASCAL 2007 dataset [23] which has been used to evaluate several image classification methods in the past. Moreover, this dataset provides bounding box annotation for many classes of objects which we use to form our OCG features.

PASCAL dataset contains 20 different object classes in total. The images are divided into three disjoint sets namely *train*, *val*, and *test*. the three data sets contain 2501, 2510, 4952 images respectively. We merge *train* and *val* sets and use it for training.

B. Implementation Details

To accumulate Shape-BoF and Color-BoF histograms, we sample image patches with the minimal size of 12x12 pixels. The neighboring patches at the same scale overlap by 50% of pixels. We use image patches of seven different sizes defined by the scale factor $\sqrt{2}$. For the baseline method (the fixed-grid strategy), we define a 3x3 image grid with cells of equal size.

The size of our shape visual vocabulary is 1000. For the color features, the size of CV1 (the first layer visual vocabulary) is equal to 128 and CV2 (the second layer vocabulary) is of size 1000. Therefore the appended feature vector that we get from each grid cell is $1000+1000 = 2000$. This implies that the length of the final feature vector in fixed-grid representation is equal to 18000.

C. OCG Configuration

Our long-term goal is to extract OCG features by considering a wide variety of possible objects in the center of the grid and combine them altogether. This would enable us to exploit both the spatial information and the semantic relations of different objects in an image. Here we simplify the problem and for each object category consider image classification problem, where we use bounding box of the target object to form our OCG. To estimate the object bounding boxes one could use a pre-trained object detector. Here we simulate results of an automatic object detector and use ground truth bounding boxes to estimate the maximum gain of OCG.

In order to get a fair conclusion when comparing fixed-grid and OCG, we exclude the center cell from the OCG cells. Therefore, we use the bounding box of objects only to put our OCG on the right position in the image. In other words, we only use the context of the object to evaluate our OCG method. Hence, OCG features are extracted in the same way as fixed-grid features except that the information from the OCG center cell is suppressed.

The final feature vector of the OCG method is the concatenation of the fixed-grid features and the features we get from the cells of OCG. Therefore, the size of the OCG feature vector is equal to $18000+16000=34000$.

If in one image, there is no object of the target class available to form the OCG, we generate a random bounding box within the boundary of the image and consider it as a simulated false positive response of an object detector as described in Section III-A. For cases with more than one target object in an image, OCG feature is calculated for each of the object instances. Alternative feature vectors are individually classified by SVM. Maximum value of the SVM outputs is then taken as the confidence value for the image.

To optimize SVM parameters, we do grid-search on *cost* and *gamma* values using libsvm toolbox [24]. Optimal parameters are found by cross-validation with 5 folds.

Table I illustrates the evaluation results of our OCG framework compared to the fixed-grid method, as well as the results of [16]. The table shows that for all of the object classes, OCG remarkably outperforms the fixed-grid approach. According to the last row of the table, OCG performs almost 15% better than the fixed-grid approach in terms of average precision over all of the 9 object classes. However, as we mentioned before, this is the maximum improvement that one could expect to get using OCG instead of the fixed-grid method. That is because the OCG is ideally localized based on ground truth annotation which simulates the ideal object detector.

In the second column of Table I, that is labeled as *Uijlings-All*, the image is represented by a BoF histogram of SIFT-like features. *Uijlings-Context-Only* is the same as *Uijlings-All* except from the fact that the object patch is removed from image before building the BoF histogram. It

(a) Car

(b) Motorbike

Figure 4. Performance of the OCG method shown with respect to the ratio of the ground truth bounding boxes that have been used when localizing the OCG. (a) is for the *Car* class and (b) is for *Motorbike*.

Table I

AVERAGE PRECISION VALUES FOR FIXED-GRID AND OCG COMPARED TO THE RESULTS OF UIJLINGS ET AL. [16]. THE ONLY DIFFERENCE BETWEEN THE *Fixed-Grid* COLUMN AND *Uijlings-All* COLUMN IS THAT THE *Fixed-Grid*, UNLIKE *Uijlings-All*, INCORPORATES SPATIAL INFORMATION INTO ITS HISTOGRAMS. THE *OCG* COLUMN CAN BE SEEN AS *Uijlings-All* COMBINED WITH *Uijlings-Context-Only* IN THE SENSE THAT THEY BOTH USE THE INFORMATION OF BOUNDING BOX ANNOTATIONS OF THE OBJECTS FROM GROUND TRUTH TO EXCLUDE THE OBJECT PATCH.

Object Category	Uijlings All	Fixed-Grid	Uijlings Context-Only	OCG
Bicycle	46.2	50.4	17.8	66.45
Car	69.0	67.2	43.1	83.79
Cat	43.7	48.8	15.5	68.11
Chair	44.9	47.7	39.0	60.81
Horse	69.2	72.9	56.4	81.94
Motorbike	49.1	56.2	25.3	75.01
Person	79.2	80.4	61.6	89.89
Sheep	28.4	34.0	15.0	54.26
TV monitor	40.9	42.1	33.7	52.17
Mean AP	52.29	55.52	34.16	70.27

means that they use ground truth annotation to cut the object out of the image and therefore use only context of the object for classification.

Even though the goal of [16] has not been to beat the state-of-the-art methods, yet the effect of incorporating spatial information is obvious from the comparison of their results to our *Fixed-Grid* results.

Aside from some minor implementation details, their experimental setting is very much similar to ours. Their evaluations is also done on PASCAL 2007. They also use BoF framework as well as SVM classification. One difference is that size of their visual vocabulary is 4096 while we use a vocabulary of size 1000, as mentioned before.

Figure 5. The two object classes *car* and *bus* share a big part of their context distribution. This image shows the first 20 high score false positives of *car* classifier using fixed-grid method. This large amount of confusion is due to the great similarity between the context as well as local appearance of the two classes.

Another difference is that they use SIFT-like features while we use HOG. However, we believe that using SIFT will further increase our performance [5]

Neglecting the aforementioned differences, there is a remarkable difference between the *Uijlings-All* column and the *Fixed-Grid* column in Table I. *Fixed-Grid* is superior because of adding spatial information into the BoF histograms.

Table I shows that for all of the object classes, except *car*, the *Fixed-Grid* is outperforming the *Uijlings-All* results. The small loss of average precision for the *Car* class turns out to be due to the confusion between the contexts of the *car* class and the *bus* class. Most of the high score false positives of the *car* class are *bus* samples (Figure 5). Uijlings et al. also mentioned in their paper that the context of *car* is a superset of the context of *bus* [16].

We want to stress once more that even though we use the bounding box annotation of objects for our evaluations,

the center cell is ignored in evaluations of OCG. Therefore, all the information available for OCG comes from the surroundings of the object which is considered as the contextual information of that object.

As a first step towards using a real object detector instead of ground truth, we make yet another experiment simulating the behavior of a detector. Results of this experiment for the *car* and *motorbike* object classes are shown in Figure 4. There are three curves for each object class in the figure. The difference between the curves is the number of random bounding boxes that we used when evaluating the method on a test image. In fact, these bounding boxes are used to form the OCGs. Therefore, for each of these random bounding boxes we get one OCG feature vector. We show the curves for 1, 10, and 20 random bounding boxes per image.

The random bounding boxes can be interpreted as false positives of the detector. The vertical axis shows average precision score. Along the horizontal axis we gradually add ground truth bounding boxes; thus the horizontal axis can be interpreted as the detection ratio of the detector.

The dashed line in Figure 4a says that if the detector produces 10 false positives per image (the red curve), and if the detector finds 30% of objects accurately, then the OCG method will start outperforming fixed-grid (cf. Table I). The more objects are detected correctly by the detector, the higher the performance of OCG will be comparing to the fixed-grid. The same situation happens for the *motorbike* class starting from 40% recall of the detector (Figure 4b).

It is worth mentioning that for the class of *car*, there are 721 objects in the database overall. Thus, 10 false positives per image, which was the case in the aforementioned scenario, means that a detector with the precision value equal to $\frac{30\% \times 721}{4898 \times 10} \approx 0.004$ is sufficient; where 4898 is the number of clear test samples. Similarly, for the *motorbike* class a detector with precision $\frac{40\% \times 222}{4941 \times 10} \approx 0.001$ is sufficient.

It may seem reasonable that if we generate one random bounding box per test image, the classification accuracy should basically be almost equal to the fixed-grid result. But, comparing the result of Table I with the leftmost point of the green curve in Figure 4 shows that fixed-grid is strictly better than OCG in such a situation. We believe that this happens because many of the images in Pascal have the target object in the center. Therefore, it is frequently the case that the central object is in the center of the image and thus is better captured by the fixed-grid.

V. CONCLUSION AND FUTURE WORK

In this paper we proposed a new method for incorporating spatial information into BoF framework. The method, which we call Object Centered Grid (OCG), works based on a central object and will model the contextual information of an image based on the distribution of surrounding of that object.

We showed that in the perfect case, when we have accurate localization of the central objects, the OCG framework will capture much more coherent and stable contextual information compared to the fixed-grid method.

In this paper our main goal was to measure the amount of information gain that we can attain by using OCG in the ideal case. The next step of our work will be to integrate the OCG idea with a real detector and to increase the performance of any object detector using OCG.

Another direction for future research is to study the possibility of fusing OCGs of different objects together. It would also be interesting to evaluate OCG on scene classification tasks using the variety of pre-trained object detectors which become accessible for the community.

ACKNOWLEDGMENT

The authors would like to thank Hossein Azizpour for the fruitful discussions and valuable comments.

REFERENCES

- [1] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, Jan. 2003.
- [2] K. P. Murphy, A. Torralba, and W. T. Freeman, "Using the forest to see the trees: a graphical model relating features, objects and scenes," in *Proc. Conf. on Neural Information Processing Systems*, Canada, 2003.
- [3] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, USA, 2009.
- [4] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, USA, 2009.
- [5] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, Apr. 2008.
- [6] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, USA, 2006.
- [7] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. European Conf. on Computer Vision*, France, Oct. 2008.
- [8] H. Maboudi, A. T. Targhi, J. O. Eklundh, and A. Pronobis, "Joint visual vocabulary for animal classification," in *Proc. International Conference on Pattern Recognition*, USA, 2008.
- [9] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, Jan. 2001.
- [10] A. Gupta and L. S. Davis, "Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers," in *Proc. European Conf. on Computer Vision*, France, 2008.

- [11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. IEEE International Conf. on Computer Vision*, Brazil, 2007.
- [12] I. Laptev, "Improving object detection with boosted histograms," *Image and Vision Computing Journal*, vol. 27, Aug. 2008.
- [13] F. Schroff, A. Criminisi, and A. Zisserman, "Single-histogram class models for image segmentation," in *Proc. Indian Conference on Computer Vision, Graphics and Image Processing*, India, 2006.
- [14] M. Marszałek, C. Schmid, H. Harzallah, and J. V. D. Weijer, "Learning object representations for visual object class recognition," oct 2007, visual Recognition Challenge workshop, in conjunction with ICCV. [Online]. Available: <http://lear.inrialpes.fr/pubs/2007/MSHV07>
- [15] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "context-based vision system for place and object recognition," in *Proc. IEEE International Conf. on Computer Vision*, Brazil, 2003.
- [16] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "What is the spatial extent of an object?" in *Proc. IEEE Conf. on Computer Vision and Patter Recognition*, USA, 2009.
- [17] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, USA, 2005.
- [18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, USA, 2005.
- [19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE International Conf. on Computer Vision*, Greece, 1999.
- [20] L. Wolf and S. Bileschi, "A critical view of context," *International Journal of Computer Vision*, vol. 62, Aug. 2006.
- [21] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM International Conference on Image and Video Retrieval*, Netherlands, 2007.
- [22] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, Sep. 2006.
- [23] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [24] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.