Objects in Context

Andrew Rabinovich
amrabino@cs.ucsd.edu

Andrea Vedaldi
avedaldi@cs.ucla.edu

Carolina Galleguillos
cgallegu@cs.ucsd.edu

Eric Wiewiora
ewiewior@cs.ucsd.edu

Serge Belongie
sjb@cs.ucsd.edu

Abstract

In the task of visual object categorization, semantic context can play the very important role of reducing ambiguity in objects' visual appearance. In this work we propose to incorporate semantic object context as a post-processing step into any off-the-shelf object categorization model. Using a conditional random field (CRF) framework, our approach maximizes object label agreement according to contextual relevance. We compare two sources of context: one learned from training data and another queried from Google Sets. The overall performance of the proposed framework is evaluated on the PASCAL and MSRC datasets. Our findings conclude that incorporating context into object categorization greatly improves categorization accuracy.

1. Introduction

Object categorization has been an active topic of research in psychology and computer vision for decades. Initially, vision scientists and psychologists formulated hypotheses about models of object categorization and recognition [8, 9, 25]. Subsequently, in the past 10 years or so, object recognition and categorization have become very popular areas of research in computer vision. With two general models emerging, generative and discriminative, the newly developed algorithms aim to adhere to the original modeling constraints proposed by vision scientists. For example, the hypothesis put forth by Biederman et al. [2] suggests five classes of relations between an object and its setting that can characterize the organization of objects into real-world scenes. These are: (i) *interposition* (objects interrupt their background), (ii) *support* (objects tend to rest on surfaces), (iii) *probability* (objects tend to be found in some context but not others), (iv) *position* (given an object is probable in a scene, it often is found in some positions and not others), and (v) *familiar size* (objects have a limited set of size relations with other objects).

Classes (i, ii, iv, and v) have been addressed fairly well in the models proposed by the computer vision community [3, 6, 24]. Class (iii), referring to the contextual interactions between objects in the scene, however, has received comparatively little attention.

Existing context based methods for object recognition and classification consider global image features to be the source of context, thus trying to capture object class specific features. In [10, 15, 26, 29], the relationship between context and object properties is based on the correlation between the statistics of low-level features across the image that contains the object, or even the whole object category.

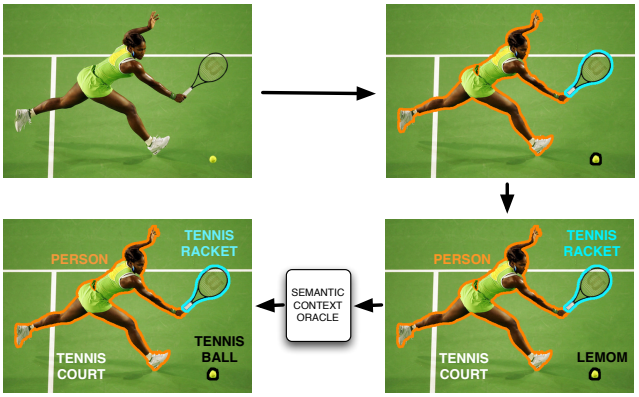


Figure 1. Idealized Context Based Object Categorization System. An original image is perfectly segmented into objects; each object is categorized; and objects' labels are refined with respect to semantic context in the image.

Semantic context [1] among objects has not been explicitly incorporated into existing object categorization models. Semantic context requires access to the referential meaning of the object [2]. In other words, when performing the task of object categorization, objects' category labels must be assigned with respect to other objects in the scene, assuming there is more than one object present. To illustrate this further, consider an example in Figure 1. In the scene of

[1] For simplicity we will use context and semantic context interchangeably from now on.

a tennis match, four objects are detected and categorized: "Tennis court", "Person", "Tennis Racket", and "Lemon". Using a categorization system without a semantic context module, these labels would be final; however, in context, one of these labels is not satisfactory. Namely, the object labeled "Lemon", with an appearance very similar to a "Tennis Ball" is probably mis-labeled, due to the ambiguity in visual appearance. By enforcing semantic contextual constraints, provided by an oracle, the label of the yellow blob changes to "Tennis Ball", as this label better fits in context with other labels more precisely.

In this work, we propose to use contextual relations between objects' labels to help satisfy semantic constraints. We extend the popular bag-of-features (BoF) model by incorporating contextual interactions between objects in the scene. In particular, we advocate using image segmentation as a pre-processing step to object categorization. Segment based representation of test images adds spatial grouping to the discriminative recognition model and provides for an intuitive representation of context based interactions between objects in the image. With object categorization in hand, a conditional random field (CRF) formulation is used as post-processing to maximize the objects' labels contextual agreement. The flow chart of our approach is shown in Figure 2.

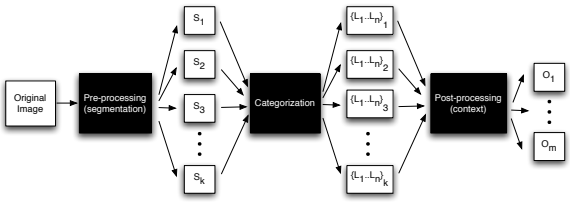


Figure 2. Object Categorization using Semantic Context. $S_1 \dots S_k$ is the set of k segments for an image drawn from multiple stable segmentations; $L_1 \dots L_n$ is a ranked list of n labels for each segment; $O_1 \dots O_m$ is a set of m objects categorizes in the original image.

The paper is organized as follows: Section 2 formalizes the theoretical framework used in this work for including context information in the object categorization task. Section 3 details the source of contextual information and its representation. In Section 4 we present experimental results. We conclude with the discussion of our approach, optimizations and future work in Section 5.

2. Object Categorization Model

Our categorization is based on the popular BoF discriminative model. As the main drawback of this type of approach is the disregard for the spatial layout of the image patches/features, we pre-process all test images with an image segmentation stage. As reported in [1], this approach significantly improves categorization accuracy of discriminative models.

2.1. Shortlist of Stable Segmentations

In an attempt to segment test images before categorization one is faced with a number of difficulties: the appropriate grouping criterion (cue selection and combination) and the number of clusters (model order). Recent advances in stability based clustering algorithms have shown promise in overcoming these problems. In this work we adopt the framework of [19] to generate a shortlist of stable segmentations.

Let us review the basics of stability based image segmentation. Cues are combined into one similarity measure using a convex combination: $W_{ij} = \sum_{f=1}^{F}(p_f \cdot C_{ij}^f)$, subject to $\sum_{f=1}^{F} p_f = 1$, where W_{ij} is the overall similarity between pixels i and j, C_{ij}^f is the similarity between the i-th and j-th pixels according to some cue f, and F is the number of cues. Since the "correct" cue combination $\vec{p}$ and the number of segments k yielding to "optimal" segmentations are unknown *a priori*, we would like to explore all possible parameter settings. However, this is not computationally viable and we adopt an efficient sampling scheme. Nonetheless, we are still left with defining the optimal segmentations, which we describe next.

Stability Based Clustering. For each choice of cue weightings $\vec{p}$ and number of segments k one obtains different segmentations of the image. Of all possible segmentations arising in this way, some subset can be considered "meaningful." Here we use stability as a heuristic to define and compute the meaningful segmentations.

For a choice of the parameters $\vec{p}$ and k, the image is segmented using Normalized Cuts [13, 22]. The segmentation is considered stable if small perturbations of the image do not yield substantial changes in the segmentation. The image is perturbed and segmented T times and the following score is evaluated: $\Phi(k, \vec{p}) = \frac{1}{n - \frac{n}{k}} \left(\sum_{i=1}^{n} \sum_{j=1}^{T} \delta_{ij} - \frac{n}{k} \right)$. Here n is the number of pixels and δ_{ij} is equal to 1 if the i-th pixel is mapped to a different segment in the j-th perturbed segmentation and zero otherwise. Thus Φ is a properly normalized[2] measure of the probability of a pixel to change label due to a perturbation of the image. Segmentations with high stability score are retained. Notice that, in general, there may exist several stable segmentations.

2.2. Bag of Features

In this work we utilize the BoF object recognition framework [5, 17] because of its popularity and simplicity. This

[2] In particular Φ ranges in $[0, 1]$ and it is not biased towards a particular value of k.

method consists of four steps: (i) images are decomposed into a collection of "features" (image patches); (ii) features are mapped to a finite vocabulary of "visual words" based on their appearance; (iii) a statistic, or *signature*, of such visual words is computed; (iv) the signatures are fed to a classifier for labeling. Here we adopt the implementation and default parameter settings provided by [27], however, a more sophisticated version of bags-of-features is likely to improve the categorization accuracy.

2.3. Integrating Bag of Features and Segmentation

We integrate segmentation into the BoF framework as follows: each segment is regarded as a stand-alone image by masking and zero padding the original image. Then the signature of the segment is computed as in regular BoF, but discarding any feature that falls entirely outside its boundary. Eventually, the image is represented by the ensemble of the signatures of its segments.

This simple idea has a number of effects: (i) by clustering features in segments we incorporate coarse spatial information; (ii) masking greatly enhances the contrast of the segment boundaries making features along the boundaries more shape-informative; (iii) computing signatures segments improves the signal-to-noise ratio.

Next we discuss how segments and their signatures are used to classify segments and whole images and to localize objects in them.

Labeling Segments. Let i be the image index, c the category index and s the segment index, so I_{ic} is the i-th training image of the c-th category. Let I be a test image and S_q its q-th segment. Let $\phi(I)$ (or $\phi(S)$) be the signature of image I (or segment S) and $\Omega(I)$ (or $\Omega(S)$) the number of features extracted in image I (or segment S).

Notice that we only segment the test images and leave the training data untouched. As such, the method does not require labeled segments for training.

Segments are classified based on a simple nearest neighbor rule. Define the un-normalized distance of the test segment S_q to class c as:

$$d(S_q, c) = \min_i d(S_q, I_{ic}) = \min_i \|\phi(S_q) - \phi(I_{ic})\|_1$$

So $d(S_q, c)$ is the minimum l_1 distance of the test segment S_q to all the training images I_{ic} of category c. We assign the segment S_q to its closest category $c_1(S_q)$:

$$c_1(S_q) = \operatorname*{argmin}_c d(S_q, c).$$

In order to combine segment labels into a unique image label it is useful to rank segments by classification reliability. To this end we introduce the following confidence measure.

Labeling Confidence. Define the *second best labeling* of segment S_q the quantity:

$$c_2(S_q) = \operatorname*{argmin}_{c \neq c_1(S_q)} d(S_q, c).$$

In order to characterize the ambiguity of the labeling $c_1(S_q)$ we compare the distance of S_q to $c_1(S_q)$ and $c_2(S_q)$. Define

$$p(c_1(S_q)|S_q) = (1-\gamma)+\gamma/C, \quad \text{where } \gamma = \frac{d(S_q, c_1(S_q))}{d(S_q, c_2(S_q))}$$

and C is the number of categories. This is the belief that S_q has class $c_1(S_q)$; for other labels, $c \neq c_1(S_q)$:

$$p(c|S_q) = \frac{1 - p(c_1(S_q)|S_q)}{C-1}.$$

Thus, $p(c|S_q)$ is a probability distribution over labels and it is uniform when $d(S_q, c_1(S_q)) \approx d(S_q, c_2(S_q))$ and peaked at $c_1(S_q)$ when $d(S_q, c_1(S_q)) \ll d(S_q, c_2(S_q))$. To reflect the importance and reliability of the segment S_q, $p(c|S_q)$ is weighted by $w(S_q) = \Omega(S_q)/\Omega(S_{\max})$, where $S_{\max}$ is the largest segment (in terms of number of features).

$$p(c|S_q) = p(c|S_q)w(S_q)$$

Localization. In many approaches to object localization, the bounding box that yields highest recognition accuracy is used to describe objects' location [14, 28]. Here we use the segment boundaries instead.

Given the labels of each segment, $c_1(S_q)$, and the overall image label, $c(I)$, we look for segments whose labels match the image label, i.e. $c(I) = c_1(S_q)$. Among these, we check for overlapping segments and we return the first k unique segment boundaries. Note that this method is not limited to BoF and could be used to add localization capabilities to other recognition methods. Given all segments S_q, we remove all overlapping segments (overlap $\geq 90\%$) and rank the remaining ones with respect to their label confidence $p(c_1(S_q)|S_q)$. The first k segment boundaries and category labels are returned.

3. Incorporating Semantic Context

To incorporate semantic context into the object categorization, we use a conditional random field (CRF) framework to promote agreement between the segment labels. CRFs have been widely used in object detection, labeling, and classification [10, 11, 15, 23]. The proposed CRF differs in two significant ways. First, we use a fully connected graph between segment labels instead of a sparse one. Second, instead of integrating the context model with the categorization model, we train the CRF on simpler problems defined on a relatively small number of segments.

Context Model. Given an image I and its segmentation $S_1, \ldots, S_k$, we wish to find segment labels $c_1, \ldots, c_k$ such

that they agree with the segment contents and are in contextual agreement with each other. We assume the labels come from a finite set $\mathcal{C}$.

We model this interaction as a probability distribution:

$$p(c_1 \dots c_k | S_1 \dots S_k) = \frac{B(c_1 \dots c_k) \prod_{i=1}^{k} A(i)}{Z(\phi, S_1 \dots S_k)}, \text{with}$$

$$A(i) = p(c_i|S_i) \text{ and } B(c_1 \dots c_k) = \exp\Big(\sum_{i,j=1}^{k} \phi(c_i, c_j) \Big),$$

where $Z(\cdot)$ is the partition function. We explicitly separate the marginal terms $p(c|S)$, which are provided by the recognition system, from the interaction potentials $\phi(\cdot)$.

To incorporate semantic context information into object categorization, namely into the CRF framework, we construct context matrices. These are symmetric, nonnegative matrices that contain the co-occurrence frequency among object labels in the training set of the database (note that both MSRC and PASCAL databases have strongly labeled training data).

Co-occurence Counts. Our first source of data for learning $\phi(\cdot)$ is a collection of multiply labeled images $I_1, \dots, I_n$. We indicate the presence or absence of label i with an indicator function l_i. The probability of some labeling is given by the model

$$p(l_1 \dots l_{|\mathcal{C}|}) = \frac{1}{Z(\phi)} \exp\Big(\sum_{i,j \in \mathcal{C}} l_i l_j \phi(i,j) \Big).$$

We wish to find a $\phi(\cdot)$ that maximizes the log likelihood of the observed label co-occurences. The likelihood of these images turns out to be a function only of the number of images, n, and a matrix of label co-occurence counts. An entry ij in this matrix counts the times an object with label i appears in a training image with an object with label j. The diagonal entries correspond to the frequency of the object in the training set. Figures 3(c) and 4(c) illustrate the structure and content of these matrices for MSRC and PASCAL training datasets respectively.

It is intractable to maximize the co-occurence likelihood directly, since we must evaluate the partition function to do this. Hence, the partition function is approximated using Monte Carlo integration [20]. Importance sampling is used where the proposal distribution assumes that the label probabilities are independent with probability equal to their observed frequency. Every time the partition function is estimated, $40,000$ points are sampled from the proposal distribution.

We use simple gradient descent to find a $\phi(\cdot)$ that approximately optimizes the data likelihood. Due to noise in estimating Z, it is hard to check for convergence; instead training is terminated when 10 iterations of gradient descent do not yield average improved likelihood over the previous 10.

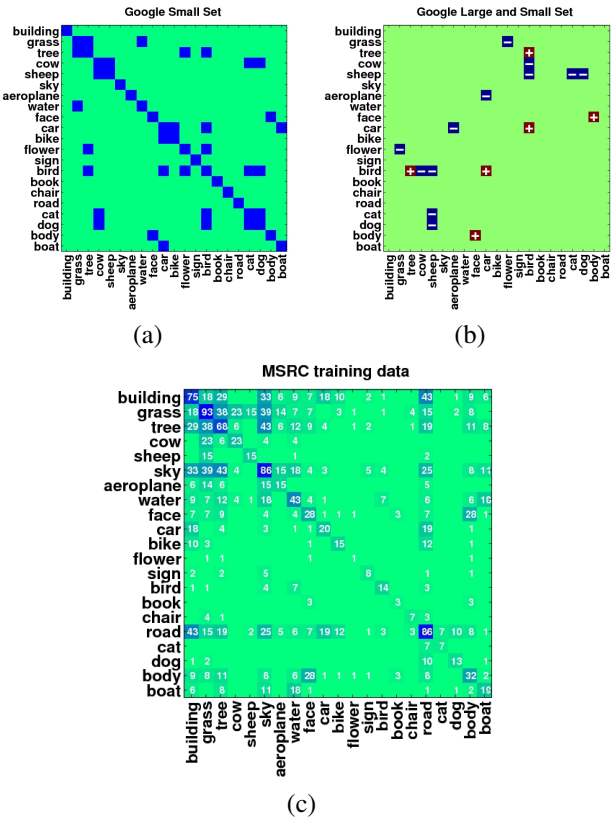


Figure 3. Context matrices for MSRC dataset. (a) Binary context matrix from GS_s. Blue pixels indicate a contextual relationship between categories. (b) Differences between small and large Google Sets context matrices. '-' signs correspond to relations present GS_s but not in GS_l; '+' correspond to relations present GS_l but not in GS_s. (c) Ground Truth, training set label co-occurence, context matrix.

Google Sets. In practice, most image databases – and images in general – do not have a training set with an equal semantic context prior and/or strongly labeled data. Thus, we would like to be able to construct $\phi(\cdot)$ from a common knowledge base, obtained from the Internet. In particular, we wish to generate contextual constraints among object categories using Google Sets[3] (GS).

Google Sets generates a list of possibly related items, or objects, from a few examples. It has been used in linguistics, cell biology and database analysis to enforce contextual constraints [7, 18, 21]. In order to obtain this information for object categorization we queried Google Sets using the labeled training data available in the MSRC and PASCAL databases. We generated a query using every category label (one example) and then matched the results against all

[3] http://labs.google.com/sets

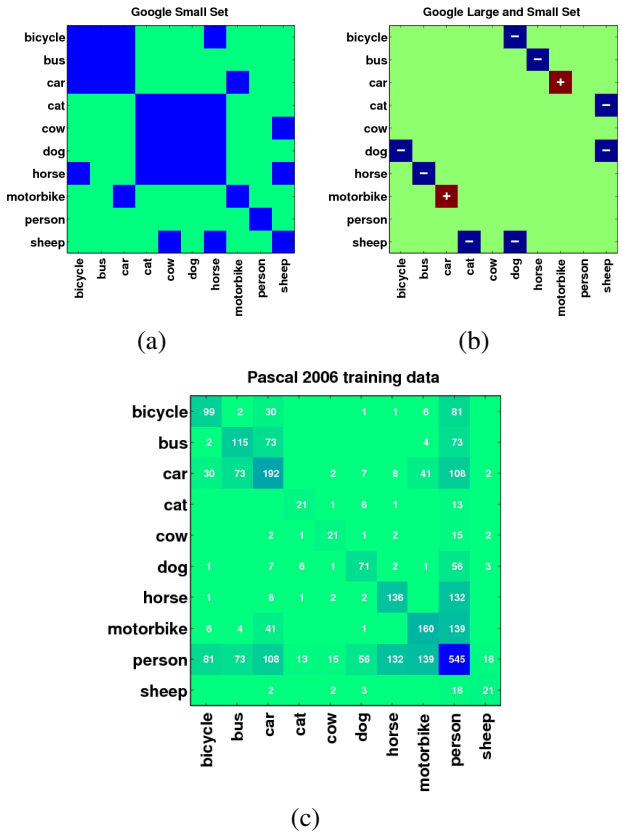


Figure 4. Context matrices for PASCAL dataset. (a) Binary context matrix from GS_s. Blue pixels indicate a contextual relationship between categories. (b) Differences between small and large Google Sets context matrices. '-' signs correspond to relations present GS_s but not in GS_l; '+' correspond to relations present GS_l but not in GS_s. (c) Ground Truth, training set label co-occurence, context matrix.

the categories present in these datasets. This task was performed for each database using the small set, GS_s, of results and the large set GS_l, which contains more than 15 results. Figures 3(a) and 4(a) show binary contexts from GS_s, for MSRC and PASCAL respectively. Intuitively, we expected $GS_S \subset GS_l$, however, $GS_s \setminus GS_l \neq \emptyset$ as shown in Figures 3(b) and 4(b). The larger set implies broader relations, thus changing the context of the set to be too general. In this work we retrieve objects labels' semantic context from GS_s.

In this case, $\phi(i,j) = \gamma$ if GS_s marks them as related, or 0 otherwise. We set $\gamma = 1$ for our experiments, though γ could be chosen using cross-validation on training data if available.

Besides Google Sets, we considered other sources of contextual information such as WordNet [4] and Word Association[4]. In the task of object categorization we found that these databases cannot offer sufficient semantic context information for the visual object categories; either due to the limited recall (in Word Association) or irrelevant interconnections (in Wordnet).

[4]http://www.wordassociation.org

4. Experimental Results and Discussion

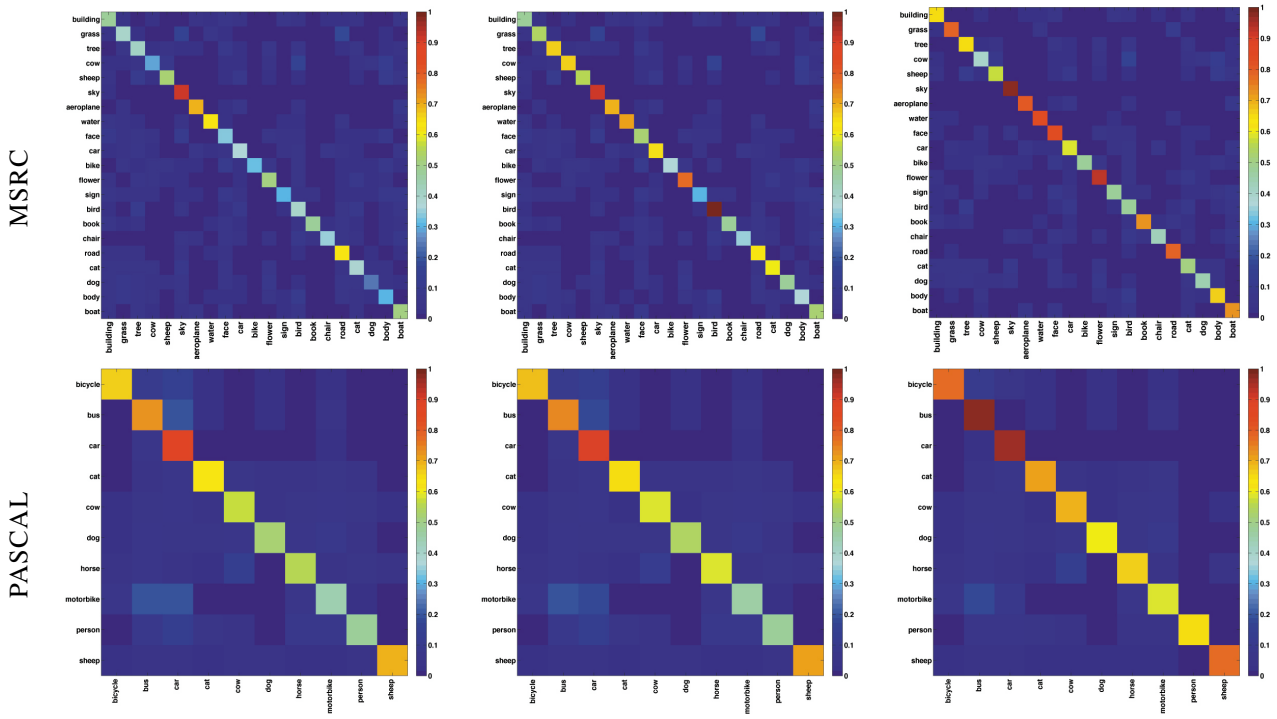
As mentioned earlier, we are interested in a relative performance change in object categorization accuracy, i.e., with and without post-processing with semantic context. In Table 1 we summarize the performance of average categorization accuracy for both the MSRC and PASCAL datasets. These results are competitive with the current state-of-the-art approaches [23, 30]. The confusion matrices, which describe the results in more details, are shown in Figure 5. For both datasets the categorization results improved considerably with inclusion of context. For the MSRC dataset, the average categorization accuracy increased by more than 10% using the semantic context provided by Google Sets, and by over 20% using the ground truth training context. In the case of PASCAL, the average categorization accuracy improved by about 2% using Google Sets, and by over 10% using the ground truth. In Figure 6 are examples where context improved object categorization. In examples 1 and 3, semantic context constraints help correct an entirely wrong appearance based labeling: bicycle – boat, and boat – cow. In examples, 2,4,5 and 6, mislabeled objects are visually similar to the ones they are confused with: boat – building, horse – dog, and dog – cow. Thus, it seems that contextual information may not only help disambiguate between visually similar objects, but also correct for erroneous appearance representation.

Clearly, context constraints can also lower or leave the categorization accuracy unchanged. As shown in Figure 7, the initially correct labels, "building" in the first image, and "grass" in the second, were relabeled incorrectly in favor of semantic context relations learned from the co-occurences in the training data. Most of such mistakes are due to the initial probability distribution over labels, $p(c|S_q)$; the feature description is not very rich as the SIFT descriptor used in this work is color-blind and segment shapes are only captured implicitly. In combining our approach with a method of strong feature description, e.g., [23], many of currently encountered errors will likely be eliminated.

	No Context	Google Sets	Using Trainig
MSRC	45.0%	58.1%	68.4%
PASCAL	61.8%	63.4%	74.2%

Table 1. Average Categorization Accuracy.

Run Time and Implementation Details. The stability based image segmentation was done with normalized cuts [22], using brightness and texture cues. A varying



(a) (b) (c)

Figure 5. Confusion matrices of average categorization accuracy for MSRC and PASCAL datasets. First row: MSRC dataset; second row: PASCAL dataset. (a) Categorization with no contextual constraints. (b) Categorization with Google Sets. context constraints. (c) Categorization with Ground Truth context constraints.

number of segments per segmentation, $k = 2, \ldots, 10$, which together results in 54 segments was considered. Implemented in MATLAB, each segmentation takes between 10-20 seconds per image, depending on the image size.

15 and 30 training images were used for the MSRC and PASCAL databases respectively. 5000 random patches at multiple scales (from 12 pixels to the image size) are extracted from each image such that larger patches are sampled less frequently (as these would be redundant). The feature appearance is represented by SIFT descriptors [12] and the visual words are obtaining by quantizing the feature space using hierarchical K-means with $K = 10$ at three levels [16]. The image signature is a histogram of such hierarchical visual words, L_1 normalized and TFxIDF reweighed [16]. In a MATLAB/C implementation, the computation of SIFT and the relevant signature, takes on average 1 second for each segment in the image. Training the classifier and constructing the vocabulary tree takes under 1 hour for 20 categories with 30 training images in each category. Classification of test images, however, is done in just a few seconds.

Training the CRF takes 3 minutes for 231 training images for MSRC and around 5 minutes for 645 images in PASCAL training dataset. Enforcing semantic constraints on a given segmentation takes between 4-7 seconds, depending on the number of segments. All the above operations were performed on a Pentium 3.2 GHz.

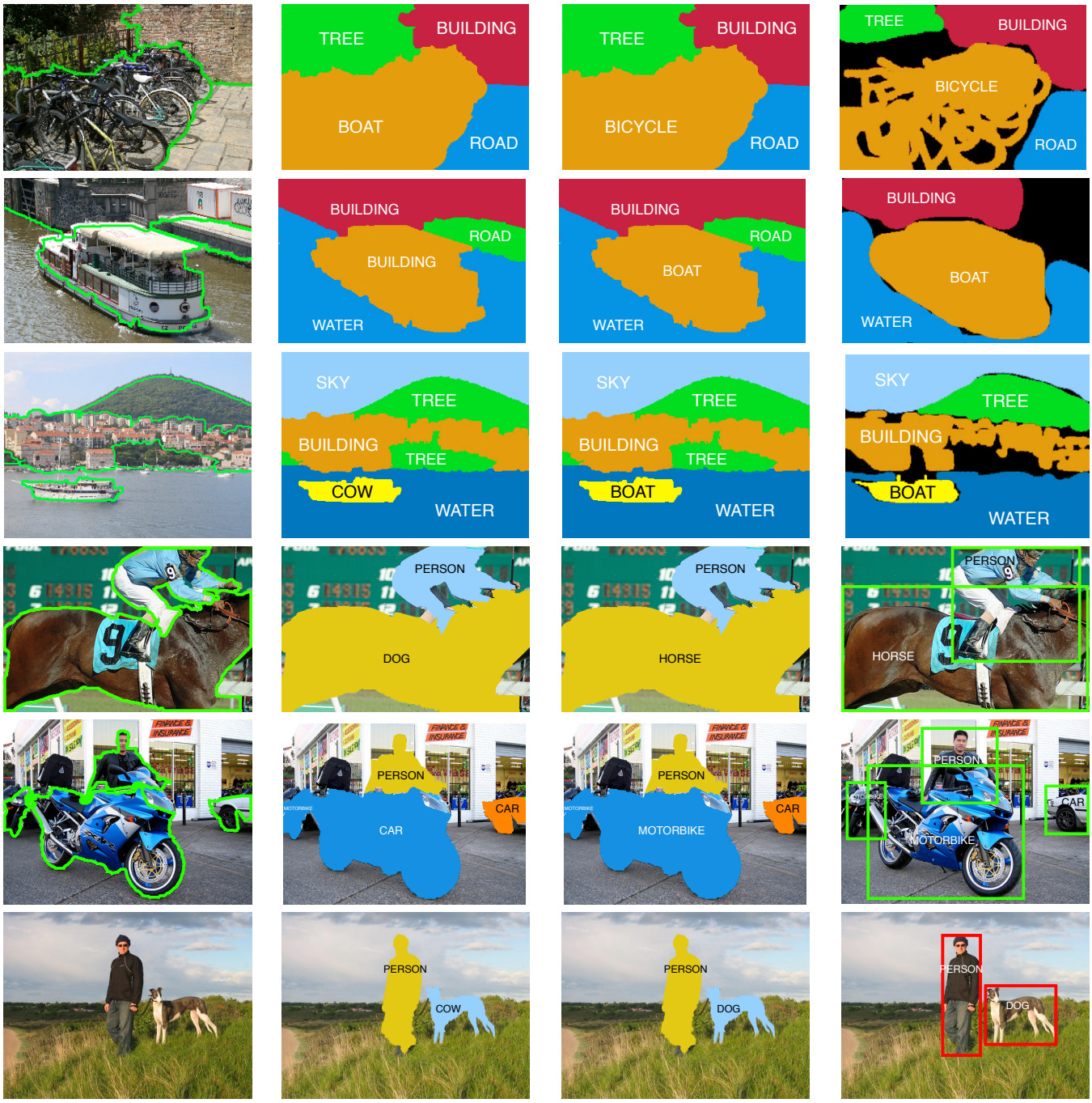
5. Conclusion

The importance of semantic context in object recognition and categorization has been discussed for many years. However, to our knowledge, there does not exist a categorization method that incorporates semantic context explicitly at the object level. In this work, we developed an approach that uses semantic context as post-processing to an off-the-shelf discriminative model for object categorization. We observed that semantic context can compensate for ambiguity in objects' visual appearance. Our approach maximizes object label agreement according to the contextual relevance.

We have studied two sources of semantic context information: the co-occurence of object labels in the training set and generic context information retrieved from Google Sets.

In addition, as pre-processing to object categorization, we advocate segmenting images into multiple stable segmentations. Using segment representations incorporates spatial groupings between image patches and provides an implicit shape description.

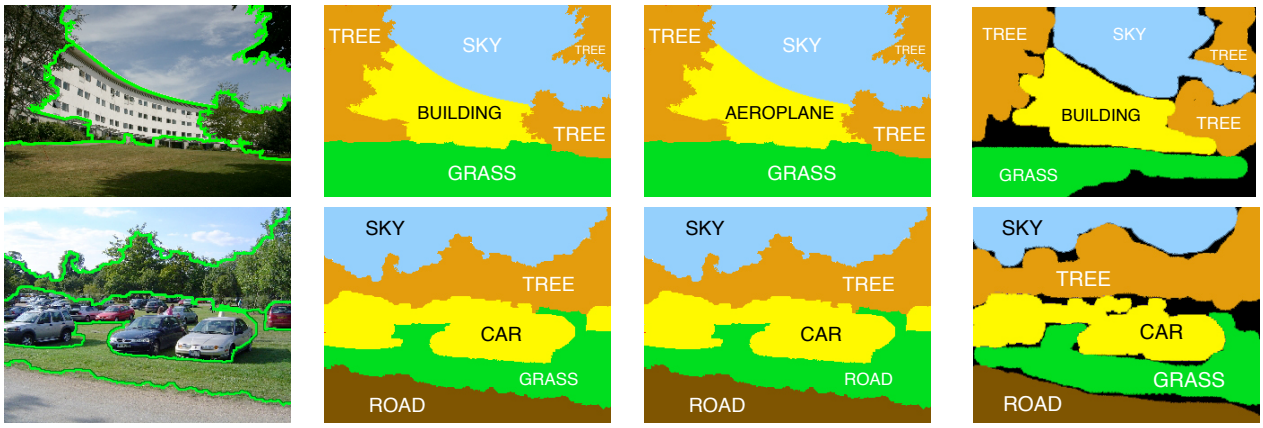
We evaluated the performance of our approach on two challenging datasets: MSRC and PASCAL. For both, the



(a) (b) (c) (d)

Figure 6. Examples of MSRC (first 3) and PASCAL (last 3) test images, where contextual constraints have improved the categorization accuracy. Results are shown in two different ways, one for each dataset. In MSRC, full segmentations of highest average categorization accuracy are shown; in PASCAL individual segments of highest categorization accuracy are shown. (a) Original Segmented Image. (b) Categorization without contextual constraints. (c) Categorization with co-occurence contextual constraints derived from the training data. (d) Ground Truth.

categorization results improved considerably with inclusion of context. For both datasets, the improvements in categorization using ground truth semantic context constraints were much higher than those of Google Sets due to the sparsity in the contextual relations provided by Google Sets. However, in considering datasets with many more cate-



(a) (b) (c) (d)

Figure 7. Examples of MSRC test images, where contextual constraints have reduced the categorization accuracy. (a) Original Segmented Image. (b) Categorization without contextual constraints. (c) Categorization with co-occurence contextual constrains derived from training data. (d) Ground Truth Categorization.

gories, we believe that context relations provided by Google Sets will be much denser and the need for strongly labeled training data will be reduced.

In our ongoing work, we are exploring alternative methods for generating semantic context relations between object categories without the use of training data. Semantic object hierarchies exist on the web, e.g., from Amazon.com, and will be utilized in this research. Finally, we are incorporating a parts-based generative model for categorization to be able to model the interactions between image segments more explicitly.

Acknowledgements

This work was funded by in part by NSF Career Grant #0448615, the Alfred P. Sloan Research Fellowship, and NSF IGERT Grant DGE-0333451.

References

[1] Anonymous. Anonymous: ICCV 2007 submission # 1404.

[2] I. Biederman, R. Mezzanotte, and J. Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–77, 1982.

[3] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

[4] C. D. Fellbaum. *WordNet : An Electronic Lexical Database*. MIT Press, 1998.

[5] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *CVPR*, 2003.

[6] M. Fink and P. Perona. Mutual boosting for contextual inference. In *NIPS*, 2004.

[7] Z. Ghahramani and K. A. Heller. Bayesian sets. In *NIPS*, 2005.

[8] A. Hanson and E. Riseman. Visions: A computer vision system for interpreting scenes. *Computer Vision Systems*, 1978.

[9] R. Haralick. Decision making in context. *PAMI*, 1983.

[10] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[11] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *ICCV*, 2003.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[13] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *ICCV*, 1999.

[14] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *CVPR*, 2006.

[15] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the tree: a graphical model relating features, objects and the scenes. In *NIPS*, 2003.

[16] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *CVPR*, 2006.

[17] E. Nowak, F. Jurie, and B. Triggs. Sampling Strategies for Bag-of-Features Image Classification. *LNICS*, 2006.

[18] J. Prager, J. Chu-Carroll, and K. Czuba. Question answering using constraint satisfaction: QA-by-dossier-with-constraints. *ACL*, 2004.

[19] A. Rabinovich, T. Lange, J. Buhmann, and S. Belongie. Model order selection and cue combination for image segmentation. In *CVPR*, 2006.

[20] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag New York, Inc., 2005.

[21] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. *JNLPBA*, 2004.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[24] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *CVPR*, 2003.

[25] T. Strat and M. Fischler. Context-based vision: Recognizing objects using information from both 2-d and 3-d imagery. *PAMI*, 1991.

[26] A. Torralba. Contextual priming for object detection. *IJCV*, 2003.

[27] A. Vedaldi. http://vision.ucla.edu/ vedaldi/code/bag/bag.html.

[28] P. Viola and M. Jones. Robust real time object detection. *IJCV*, 2002.

[29] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 2006.

[30] H. Zhang, A. C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.