Object Categorization using Co-Occurrence, Location and Appearance

Carolina Galleguillos Andrew Rabinovich Serge Belongie
Department of Computer Science and Engineering
University of California, San Diego
{cgallegu,amrabino,sjb}@cs.ucsd.edu

Abstract

In this work we introduce a novel approach to object categorization that incorporates two types of context – co-occurrence and relative location – with local appearance-based features. Our approach, named CoLA (for Co-occurrence, Location and Appearance), uses a conditional random field (CRF) to maximize object label agreement according to both semantic and spatial relevance. We model relative location between objects using simple pairwise features. By vector quantizing this feature space, we learn a small set of prototypical spatial relationships directly from the data. We evaluate our results on two challenging datasets: PASCAL 2007 and MSRC. The results show that combining co-occurrence and spatial context improves accuracy in as many as half of the categories compared to using co-occurrence alone.

1. Introduction

Real world scenes often exhibit a coherent composition of objects, both in terms of relative spatial arrangement and co-occurrence probability. This type of knowledge can be a strong cue for disambiguating object labels in the face of clutter, noise and variation in pose and illumination. Information about typical configurations of objects in a scene has been studied in psychology and computer vision for years, in order to understand its effects in visual search, localization and recognition performance [1, 2, 3, 11, 17, 28]. Bar *et al*. [1] examined the consequences of pairwise spatial relations between objects that typically co-occur in the same scene on human performance in recognition tasks. Their results suggested that (i) the presence of objects that have a unique interpretation improve the recognition of ambiguous objects in the scene, and (ii) proper spatial relations among objects decreases error rates in the recognition of individual objects.

Some recently developed computational models have appealed to observation (i) in order to identify ambiguous objects in a scene. Torralba *et al*. [25] suggested a low level representation of an image called the "Gist" as a contextual prior for object recognition. Along these lines, other approaches have also considered global image features as a source of context; either by using the correlation of low level features across images that contain the object or across the category [8, 14, 23, 27, 29]. Most recent methods incorporate co-ocurrence of high level features to enforce contextual constraints [4, 18]; information from an external knowledge base may be used instead of learned co-occurences [18].

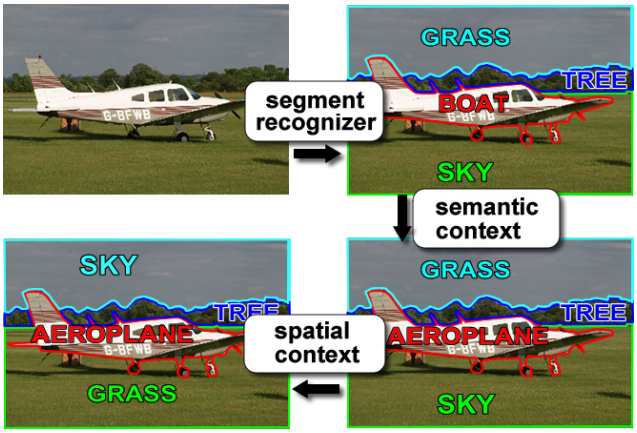


Figure 1. Illustration of an idealized object categorization system incorporating semantic and spatial context. First, the input image is segmented, and each segment is labeled by the recognizer. Next, semantic context is used to correct some of the labels based on object co-occurrence. Finally, spatial context is used to provide further disambiguation based on relative object locations.

With respect to observation (ii), several object recognition models have been proposed that consider spatial relationships. Spatial context has been incorporated from inter-pixel statistics [8, 23, 26, 29] and from pairwise relations between regions in images [9, 24]. To our knowledge, however, such models have not yet incorporated the use of explicit spatial context between *objects* in a scene. While the work of [18] employs *semantic context*[1] to improve recogni-

[1]Throughout the paper we will use semantic context and co-occurence interchangeably.

tion accuracy by maximizing label agreement of the objects in a scene with respect to co-occurrence, it does not place constraints on the relative locations of the objects.

As an illustration of this idea, consider the flow chart in Figure 1. An input image containing an aeroplane, trees, sky and grass (top left) is first processed through a segmentation-based object recognition engine. The recognizer outputs an ordered shortlist of possible object labels; only the best match is shown for each segment (top right). Without appealing to context, several mistakes are evident. Semantic context in the form of probable object co-occurrence allows one to correct the label of the aeroplane, but leaves the labels of the sky and grass incorrect (bottom right). Finally, spatial context asserts that sky is more likely to appear above grass than *vice versa* (bottom left).

Our primary contribution in this paper is a new method of object categorization that incorporates both of the above types of context into a unified framework. Our approach, named CoLA (for Co-occurrence, Location and Appearance), uses a conditional random field (CRF) formulation in order to maximize contextual constraints over the object labels. Co-occurrence and spatial context are learned simultaneously from the training data in an unsupervised manner, and models for spatial relationships between objects are discovered, rather than defined *a priori* as in [9, 24].

Our approach leverages the use of multiple stable segmentations as pre-processing step [13, 18, 21]. This representation provides a natural spatial grouping of pixels inside candidate object regions, thereby leading to improved performance of simple recognition approaches such as Bag of Features (BoF), and additionally, it readily lends itself to object-based contextual reasoning [19]. Additionally, multiple stable segmentations are a convenient substrate for object-based contextual reasoning.

The remainder of this paper is organized as follows. Section 2 describes our proposed model for learning spatial relationships between objects. In Section 3 we present and formalize the CoLA object categorization framework. Section 4 presents experimental results for two challenging databases, PASCAL 2007 and MSRC. Finally, in Section 5, we present our conclusions and discuss future work.

2. Learning Spatial Context

Biederman *et al*. [3] proposed that physical and semantic changes in a coherent scene interfere with and cause delays in object recognition. Conversely, object recognition can be facilitated by the use of relationships that support the definition of a coherent scene.

In the area of object recognition and scene understanding, several works have incorporated the use of spatial relationships as a source of context. The work of Singhal *et al*. [24] combines probabilistic spatial context models and material detectors for scene understanding. These models are based on pre-defined pixel level relationships between image regions, where spatial context information is represented as a binary feature of each specified relationship. Kumar and Hebert [9] model interactions among pixels, regions and objects using a hierarchical CRF. In their approach, the computed regions and objects are a result of the CRF itself. Although it is possible to capture a variety of different low level pixel groupings in the first level of their hierarchy, the authors only consider a single equilibrium configuration and propagate it (along with its uncertainty) to the level of regions and objects.

In contrast, our approach employs a decoupled segmentation stage that extracts a shortlist of stable (and possibly overlapping) segments as input to a subsequent context based reasoning stage. As a result, the latter stage – also CRF-based – has at its disposal a variety of shortlists of possible objects and labels over which to perform inference based on co-occurrence and spatial relationships. These relationships, which in our case are unknown *a priori*, characterize the nature of object interaction in real world images and reveal important information to disambiguate object identity.

Our sources of information for learning spatial configurations on pairs of objects are the MSRC and PASCAL training databases. In particular, these datasets provide us a collection of multiply labeled images $I_1, \ldots, I_n$, each containing at least two objects belonging to different categories, $c_i, c_j \in \mathcal{C}$ s.t. $i \neq j$; an object i is labeled by a bounding box or pixel mask β_i. We define the following simple pairwise feature to capture a specific object configuration as a three dimensional spatial context descriptor:

$$F_{ij} = (\mu_{ij}, O_{ij}, O_{ji})^\top \quad \forall i, j \in \mathcal{C},\ i \neq j, \tag{1}$$

$$O_{ij} = \frac{\beta_i \,/ \beta_j}{\beta_i} \text{ and } \mu_{ij} = \mu_{yi} - \mu_{yj} \tag{2}$$

where μ_{ij} is the difference between the y component of the centroids (in normalized coordinates) of the objects labeled c_i and c_j, and O_{ij} is the overlap percentage of the object with label c_j with respect to the object with label c_i. We omit the x component of the centroid since relative horizontal position does not carry any discriminative information for the objects in PASCAL or MSRC.

In order to capture the prevalent spatial arrangements among objects in the databases, we vector quantize the feature space into 4 groups. Choosing a small number of groups translates into simpler relations that can explain interactions that are well represented across many object pairs and scenes. We used the ground truth segmented regions and bounding box labels from MSRC and PASCAL 2007, respectively, to compute the spatial context descriptors. A

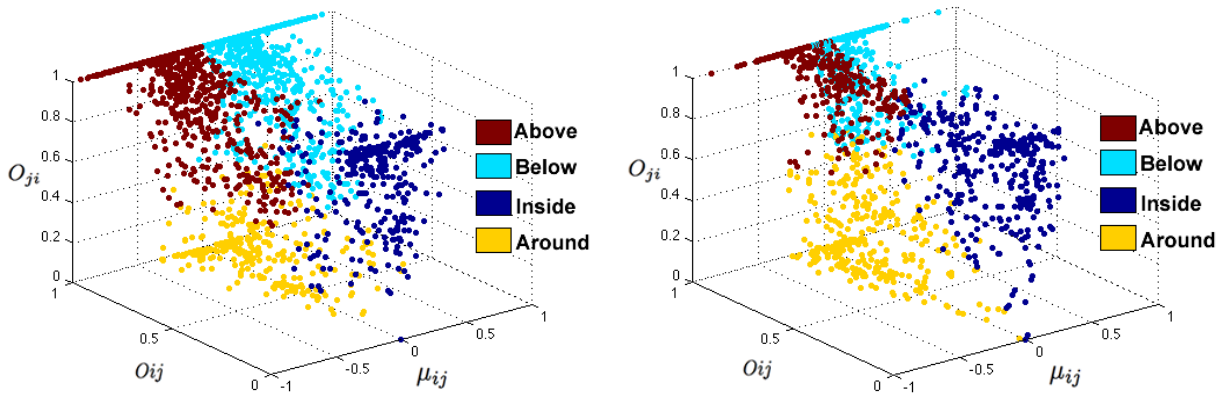


Figure 2. Four different groups represent four different spatial relationships: *above*, *below*, *inside* and *around*. The axes O_{ij}, O_{ji} and μ_{ij} are defined in Equation 2. (a) For MSRC we observe many more pairwise relationships that belong to vertical arrangements. (b) For PASCAL 2007 we observe comparatively more pairwise relationships that belong to overlapping arrangements. *Please view in color.*

closer look at the resultant clusters, shown in Figure 2, suggests the pairwise relationships *above*, *below*, *inside* and *around*, illustrated for an example image in Figure 3 containing *grass*, *water* and *cow*. Learning the relationships between pairs of objects, rather than defining them *a priori*, yields a more generic and robust description of spatial interactions among objects.

The distributions we observe in Figure 2 have comparable overall shapes, and the clusters representing the spatial relations are found in similar locations in the feature space. In the case of MSRC, the *above* and *below* relationships are predominant, as many objects remain in vertically consistent locations relative to other objects (e.g., sky, water, grass). In contrast, PASCAL's biggest clusters correspond to the spatial relationships *inside* and *around*, since most of these objects are found interposed with respect to one another. Also, as PASCAL object labels are specified by bounding boxes, rather than pixel-resolution ground truth masks, this results in larger average overlap values.

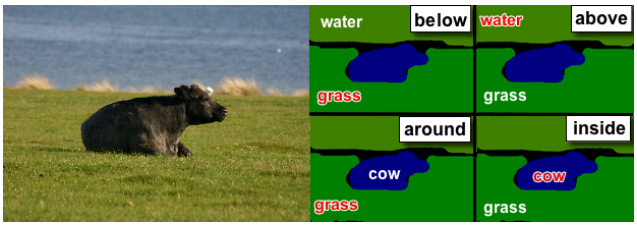


Figure 3. Illustration of four basic spatial relationships that exist among objects within an MSRC image. Labels in red indicate the object that possesses the relationship with respect to the object with the white label, e.g, the grass, in red, is below water, in white. *Please view in color.*

3. Contextual Object Categorization Model

In this section we present the details of our proposed model. At a high level, we begin by computing multiple stable segmentations for the input image, resulting in a large collection of segments. Each segment is considered as an individual image and is used as input to a BoF model for recognition. Each segment is assigned a list of candidate labels, ordered by confidence. The segments are modeled as nodes of a CRF, where location and object co-occurence constraints are imposed. Finally, based on local appearance, contextual agreement and spatial arrangements, each segment receives a category label. A flow diagram of this model is shown in Figure 4, and the details are provided next.

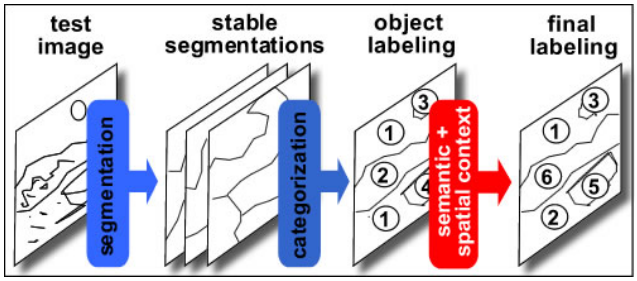


Figure 4. Object categorization using semantic and spatial context. Semantic and spatial information are unified in the same level in a conditional random field in order to constrain the location and co-ocurrence of objects in the image scene.

3.1. Appearance

BoF is a widely used discriminative model for recognition [7, 16]. Empirically, it has been shown to be rather powerful, however, it is highly sensitive to clutter, because

no distinction between object and background is made. In the raw formulation of BoF, there is no regard for spatial arrangement among pixels, patches, or features. A number of methods have been proposed to incorporate spatial information into BoF [10, 13, 18, 19]. In this work we adopt the approach of [19], which demonstrates an improvement in categorization accuracy using multiple stable segmentations.

We integrate segmentation into the BoF framework as follows. Each segment is regarded as a individual image by masking and zero padding the original image. As in regular BoF, the signature of the segment is computed, but features that fall entirely outside of segment boundary are discarded. 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) the masking step generally enhances the contrast of the segment boundaries, thereby making features along the boundaries more shape-informative; (iii) computing signatures on segments improves the signal-to-noise ratio. More details of combining stable segmentations with BoF can be found in [18].

3.2. Location and Co-Ocurrences

To incorporate spatial and semantic context into the recognition system, we use a CRF to learn the conditional distribution over the class labeling given an image segmentation. Previous works in object recognition, classification and labeling have benefited from CRFs [8, 9, 14, 23]. Our CRF formulation uses a fully connected graph between segment labels instead of a sparse one, which yields a much simpler training problem, since the random field is defined over a relatively small number of segments rather than a huge number of raw pixels or small patches.

Context Model. Given an image I, its corresponding segments $S_1, \ldots, S_k$, and probabilistic per-segment labels $p(c_i|S_i)$ (as in [18]), we wish to find segment labels $c_1, \ldots, c_k \in \mathcal{C}$ such that all agree with the segments' content and are in contextual agreement with one other.

We model this interaction as a probability distribution:

$$p(c_1 \ldots c_k|S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} p(c_i|S_i)}{Z(\phi_0, \ldots \phi_r, S_1 \ldots S_k)},$$

$$\text{with } B(c_1 \ldots c_k) = \exp\Big(\sum_{i,j=1}^{k} \sum_{r=0}^{q} \alpha_r \phi_r(c_i, c_j) \Big),$$

where $Z(\cdot)$ is the partition function, α_r a parameter estimated from training data and q is the number of pairwise spatial relations. We explicitly separate the marginal terms $p(c|S)$, which are provided by the recognition system, from the interaction potentials $\phi_r(\cdot)$. To incorporate both semantic and spatial context information into the CRF framework, we construct context matrices, described next.

Location. Spatial context is captured by frequency matrices for each of the four pairwise relationships (*above*, *below*, *inside* and *around*). The matrices contain the occurrence among objects labels in the four different configurations, as they appear in the training data. An entry (i, j) in matrix $\phi_r(c_i, c_j)$, with $r = 1, \ldots, 4$, counts the number of times an object with label i appears with an object label j for a given relationship r.

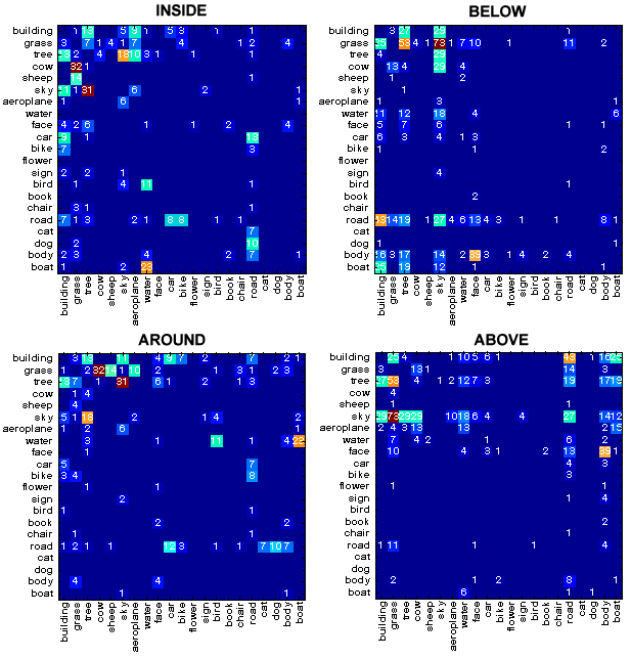


Figure 5. Frequency matrix for spatial relationships *above*, *below*, *inside* and *around* for MSRC database. Each entry (i, j) in a matrix counts the number times an object with label i appears in a training image with an object with label j according to a given pairwise relationship.

Figures 5 and 6 illustrate the counts over the four different relationships for MSRC and PASCAL. It is worth noting that MSRC matrices exhibit more uniform interactions between objects, while matrices of PASCAL single out categories of very high activity (e.g., *person*).

Co-occurrence Counts. While the occurrence of category labels are captured by the spatial context matrices above, the appearance frequency – a parameter required for the CRF – is not captured explicitly, since these matrices are hollow. Using the existing spatial context matrices, object appearance frequency can be computed as row sums of all for matrices. Finally, the sum of all four matrices, including the row sums, will result in a marginal (i.e., without regard for location) co-occurrence matrix, equivalent to those pre-

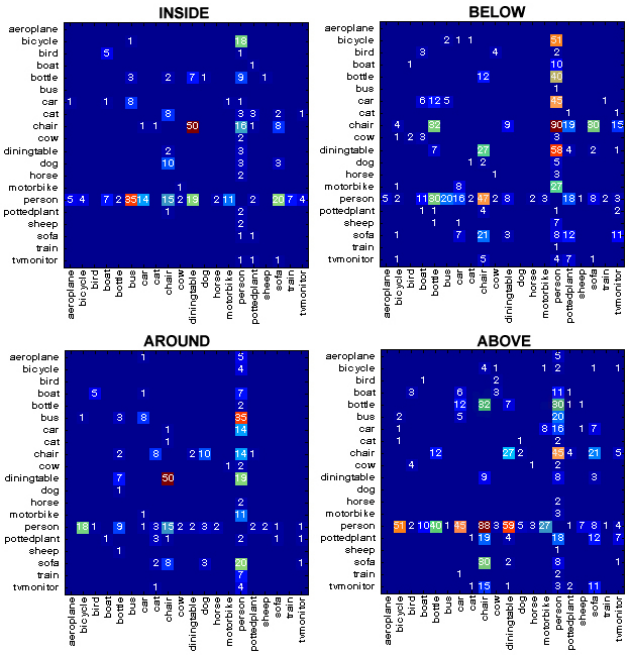


Figure 6. Frequency matrix for spatial relationships *above*, *below*, *inside* and *around* for PASCAL database. Each entry (i, j) in a matrix counts the times an object with label i appears in a training image with an object with label j given their pairwise relationship.

sented in [18]. An entry (i, j) in the semantic context matrix counts the number of 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:

$$\phi_0(c_i, c_j) = \phi'(c_i, c_j) + \sum_{k=1}^{|\mathcal{C}|} \phi'(c_i, c_k)$$

where $\phi'(\cdot) = \sum_{r=1}^{q} \phi_r(c_i, c_j)$. Therefore the probability of some labeling is given by the model

$$p(l_1 \dots l_{|\mathcal{C}|}) = \frac{1}{Z(\phi)} \exp\Big(\sum_{i,j \in \mathcal{C}} \sum_{r=0}^{q} l_i l_j \cdot \alpha_r \cdot \phi_r(c_i, c_j) \Big),$$

with l_i indicating the presence or absence of label i. We wish to find a $\phi(\cdot)$ that maximizes the log likelihood of the observed label co-occurences. Since we must evaluate the partition function, maximizing the co-occurence likelihood directly is intractable. Therefore we approximate the partition function 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. The likelihood of these images turns out to be a function only of the number of images, n, and the co-occurence matrices $\phi_r(c_i, c_j)$.

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.

4. Experimental Results

To evaluate categorization accuracy of the proposed model and the relative importance of spatial context in this task, we consider MSRC and PASCAL 2007 datasets. Table 1 summarizes the performance of average categorization per category.

Categories	Semantic Context [18]	CoLA
building	0.85	**0.91**
grass	0.94	**0.95**
tree	0.78	**0.80**
cow	0.36	**0.41**
sheep	0.55	0.55
sky	0.89	**0.97**
aeroplane	0.73	0.73
water	0.95	0.95
face	0.80	**0.81**
car	0.57	0.57
bike	0.59	**0.60**
flower	0.65	0.65
sign	0.54	0.54
bird	0.54	*0.52*
book	0.56	0.56
chair	0.42	0.42
road	0.94	**0.96**
cat	0.42	0.42
dog	0.46	0.46
body	0.75	**0.77**
boat	0.76	**0.81**

Categories	Semantic Context [18]	CoLA
aeroplane	0.63	0.63
bicycle	0.22	0.22
bird	0.18	*0.14*
boat	0.28	**0.42**
bottle	0.43	0.43
bus	0.46	**0.50**
car	0.62	0.62
cat	0.32	0.32
chair	0.37	0.37
cow	0.19	0.19
dining table	0.30	0.30
dog	0.32	*0.29*
horse	0.12	**0.15**
motorbike	0.31	0.31
person	0.43	0.43
potted plant	0.33	0.33
sheep	0.41	0.41
sofa	0.37	0.37
train	0.29	0.29
tv monitor	0.62	0.62

Table 1. Comparison of recognition accuracy between the models for MSRC and PASCAL categories. Results in **bold** indicate an increase in performance by our model. A decrease in performance is shown in *italics*.

These results outperform current state-of-the-art approaches [6, 23] and the average categorization per database is 68.38% for MSRC and 36.7% for PASCAL. What is of more interest to us, however, is the per category accuracy as a function of the type of context used. Specifically, we notice that around half of the 21 categories in MSRC benefit from using spatial context: an increase from 1%-8% in recognition accuracy. For the rest of the categories, in turn, spatial context did not harm the performance, except for a small decrease in accuracy on category *bird*.

In the PASCAL database, the availability of spatial context data is less uniform across categories. An improvement is seen in only three categories, though in one case (for category *boat*) this increase was rather high (14%). As with

MSRC, the other categories are largely unaffected by spatial context, and only one category (*bird*) suffers from reduced accuracy.

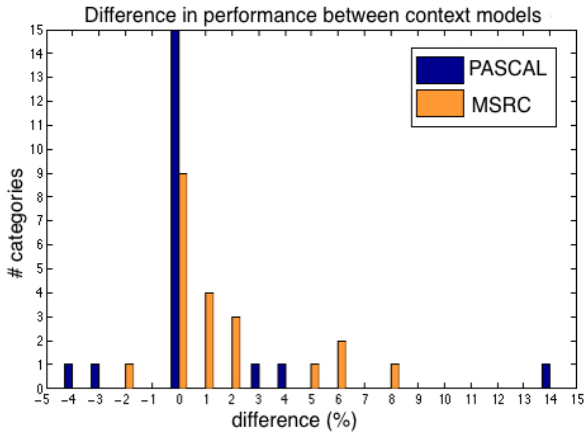


Figure 7. Difference in performance between semantic and semantic+spatial framework for MSRC and PASCAL databases.

Figure 7 summarizes the relative improvement of categorization accuracy with the inclusion of spatial context into the recognition model. Very few categories' accuracies are worsened by spatial context; most are either unchanged or improved. Some examples of affected categories are shown in Figures 8 and 9.

Run Time and Implementation Details. Stability based image segmentation was implemented with normalized cuts [5, 22], using brightness and texture cues. We considered 9 segmentations per test image, where the number of segments per segmentation ranges from $k = 2, \ldots, 10$. The computation time for each segmentation is between 10-20 seconds per image. As the individual segmentations are independent of one another, we computed them all in parallel on a cluster. As a result, a computation of all stable segmentations per image requires about 10 minutes.

15 and 30 training images were used for the MSRC and PASCAL databases respectively. 5000 random patches at multiple scales (from 12 pixels up to the image size) are extracted from each image. The feature appearance is represented by SIFT descriptors [12] and the visual words are obtained by quantizing the feature space using hierarchical K-means with $K = 10$ at three levels [15]. The image signature is a histogram of such hierarchical visual words, L_1 normalized and TFxIDF re-weighed [15]. The computation of SIFT and the relevant signature, implemented in C, takes on average 1.5 seconds per segment. Training and constructing the vocabulary tree requires less than 40 minutes for 20 categories with 30 training images in each category, in the case of PASCAL. Classification of test images is done in just a few seconds. Training the CRF takes 3 minutes for 315 training images for MSRC and 5 minutes for 600 images in PASCAL training dataset. Enforcing semantic and spatial 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 and Future Work

We have presented a novel framework for object categorization, named CoLA, that uses a CRF to maximize object label agreement in the scene according to spatial and co-ocurrence constraints. We express relative location between objects using a simple pairwise feature. By vector quantizing the feature space, we learn four different spatial relationships corresponding to *above*, *below*, *inside* and *around*. Incorporating spatial relationships into the categorization model of [18] improves recognition accuracy in many categories and gives further insight into the challenges in object categorization.

Clearly, spatial information that captures the relative object location in an image is a strong visual cue. However, unlike simple co-occurrence relationships, which can be learned from auxiliary sources such as Google Sets, spatial context must be learned directly from the training data. As our experiments have shown, spatial context learned from both MSRC and PASCAL datasets is highly nonuniform. In particular, spatial interactions among different categories are rather sparse, and many valid objects that appear in the scenes are simply considered clutter, and thereby cannot contribute contextual value. With the continued introduction of publicly available datasets possessing increasingly detailed annotations over larger numbers of categories, our proposed system is designed to scale favorably: stronger semantic and spatial context will provide more avenues for improving categorization accuracy.

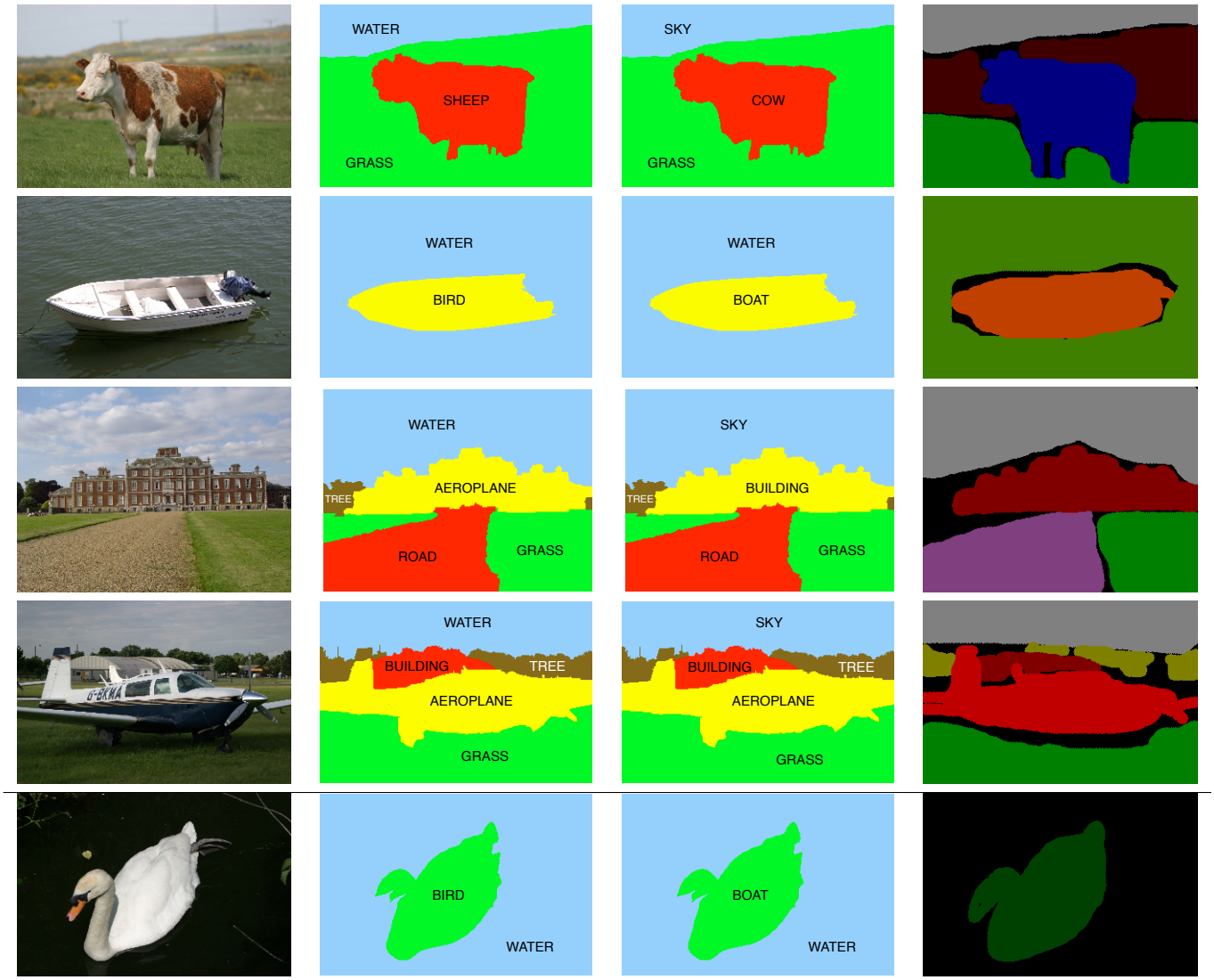
In our ongoing work, we aim to integrate the proposed model into systems for image retrieval, image annotation and event classification. In addition, we are exploring richer descriptions of object shape and alternative characterizations of spatial relationships with finer granularity.

Acknowledgements: This work was funded in part by NSF Career Grant #0448615, the Alfred P. Sloan Research Fellowship, NSF IGERT Grant DGE-0333451 and the Google Research Award.This is a corrected version of the paper appeared in CVPR'08.

References

[1] M. Bar and S. Ullman. Spatial context in recognition. *Perception. 25:343-352.*, 1993.

[2] I. Biederman. Perceiving real-world scenes. *Science*, 177(7):77–80, 1972.



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

Figure 8. Example results from the MSRC database. Spatial constraints have improved (first four rows) and worsened (last row) the categorization accuracy. Full segmentations of highest average categorization accuracy are shown. (a) Original image. (b) Categorization with co-occurrence contextual constraints [18]. (c) Categorization with spatial and co-occurence contextual constraints. (d) Ground Truth.

[3] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, April 1982.

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

[5] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR*, 2005.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

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

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

[9] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. *ICCV*, 2005.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[11] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In *CVPR*, 1997.

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

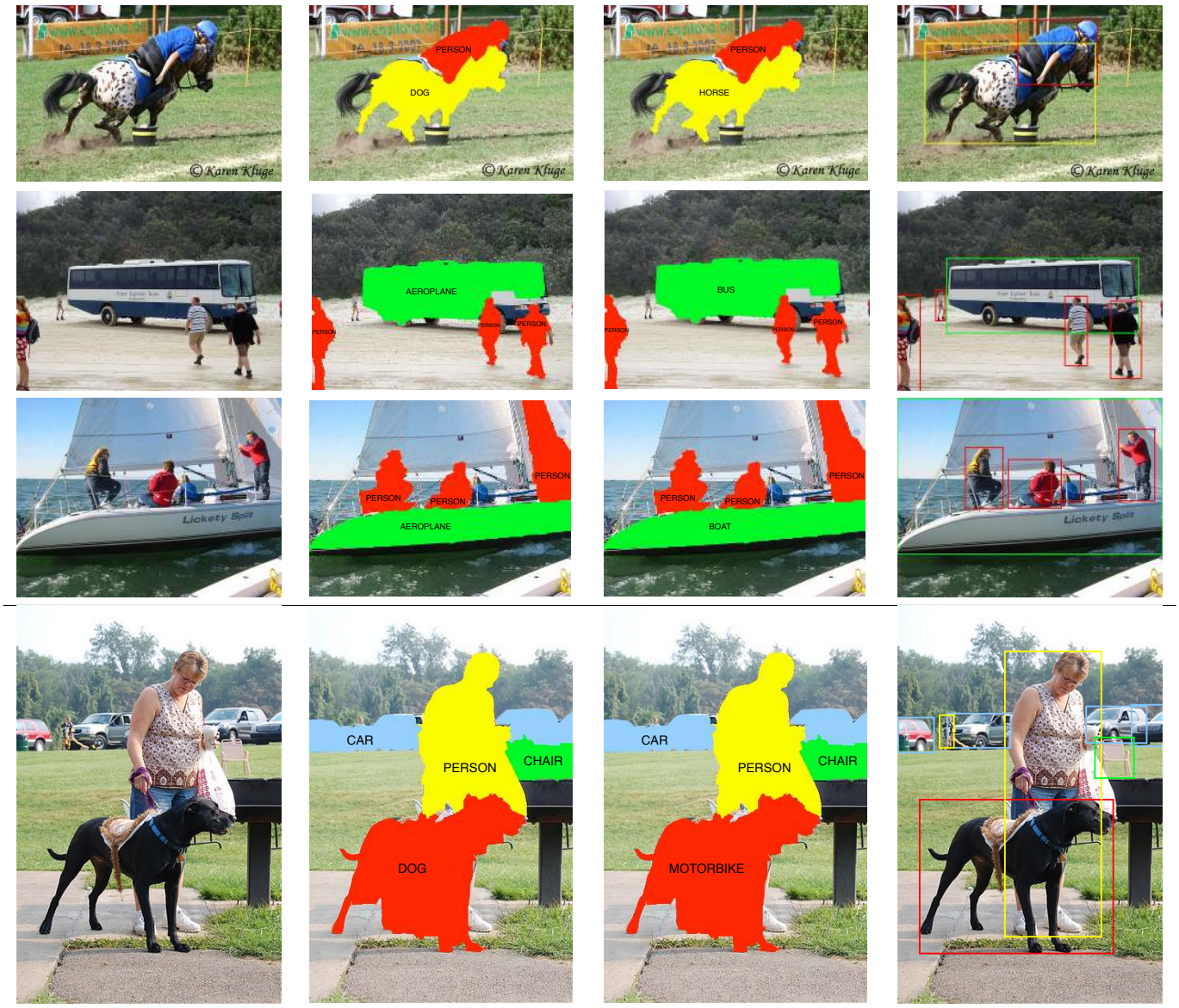
[13] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007.

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

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

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

[17] S. E. Palmer. The effects of contextual scenes on the identification of objects. *Memory and Cognition*, 1975.



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

Figure 9. Example results from the PASCAL 07 database. Spatial constraints have improved (first four rows) and worsened (last row) the categorization accuracy. Individual segments of highest categorization accuracy are shown. (a) Original image. (b) Categorization with co-occurrence contextual constraints [18]. (c) Categorization with spatial and co-occurence contextual constraints. (d) Ground Truth.

[18] A. Rabinovich, A.Vedaldi, C. Galleguillos, E. Wiewora, and S. Belongie. Objects in context. *ICCV*, 2007.

[19] A. Rabinovich, A. Vedaldi, and S. Belongie. Does image segmentation improve object categorization? *UCSD Technical Report cs2007-0908*, 2007.

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

[21] V. Roth and B. Ommer. Exploiting low-level segmentation for object recognition. *DAGM*, 2006.

[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 for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *IJCV*, pages 1–22, 2007.

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

[25] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[26] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *CVPR*, 2003.

[27] J. Verbeek and B. Triggs. Scene segmentation with CRFs learned from partially labeled images. In *NIPS*, 2007.

[28] L. Wixson and D. Ballard. Using intermediate objects to improve the efficiency of visual search. *IJCV*, 12(2):209–230, 1994.

[29] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 69(2):251–261, 2006.