Learning Object Relationships via Graph-based Context Model

Heesoo Myeong   Ju Yong Chang*   Kyoung Mu Lee

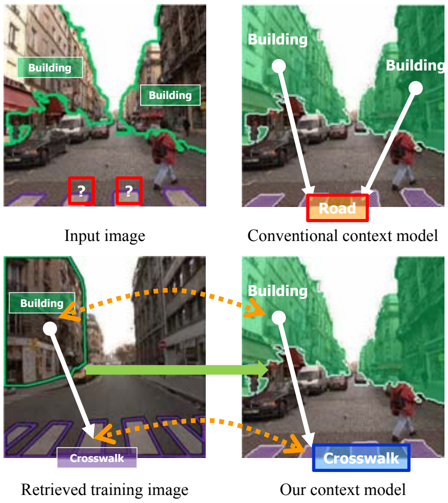
Department of EECS, ASRI, Seoul National University, 151-742, Seoul, Korea
{heesoo.myeong, juyong.chang}@gmail.com, kyoungmu@snu.ac.kr
http://cv.snu.ac.kr

Abstract

In this paper, we propose a novel framework for modeling image-dependent contextual relationships using graph-based context model. This approach enables us to selectively utilize the contextual relationships suitable for an input query image. We introduce a context link view of contextual knowledge, where the relationship between a pair of annotated regions is represented as a context link on a similarity graph of regions. Link analysis techniques are used to estimate the pairwise context scores of all pairs of unlabeled regions in the input image. Our system integrates the learned context scores into a Markov Random Field (MRF) framework in the form of pairwise cost and infers the semantic segmentation result by MRF optimization. Experimental results on object class segmentation show that the proposed graph-based context model outperforms the current state-of-the-art methods.

1. Introduction

Scene understanding is one of the core problems in computer vision. Recent works [5, 7, 9, 10, 11, 14, 21, 22, 29] have shown that employing contextual information is extremely helpful for resolving this problem. There are various sources of context including scene [10, 29], semantic [21, 22], scale [7, 21], and spatial relation [6, 8]. Recently, many researchers have highlighted the importance of pairwise relationships between objects [6, 8, 11, 21, 22]. This relationship is commonly represented by high-level statistics such as the object class co-occurrence which captures semantic context between object classes. For example, building and road are likely to co-occur in an image. To incorporate object relationships, traditional approaches often model such relations as local interactions between pixels or regions. To produce the final labeling result, the obtained object relationships are combined with pre-learned unary potentials which are usually learned based on the visual features of the objects. This scenario, separately learning contextual relationships and visual appearance, has been successfully used to solve the scene understanding problem.



Input image   Conventional context model

Retrieved training image   Our context model

Figure 1. Comparison of our context model to a conventional context model based on co-occurrence statistics. We appropriately establish the object relationship depend on the visual appearance as well as the contextual relation from the matched similar scene.

However, this system tends to prefer frequently appeared objects to enforce object label agreement according to semantic relevance. For example, consider the example illustrated in Figure 1, where the ground truth label of the unknown regions is *crosswalk*. Notice that the regions labeled as *building* enforce the unknown regions to be labeled as *road* because building and road are more strongly correlated than building and crosswalk. Furthermore, as pointed out in [15], conventional context models are not invariant to the number of pixels/regions that an object occupies, which makes the small objects likely to be eliminated. Our key idea is to utilize context relationships adaptively according

*Currently at Electronics and Telecommunications Research Institute in Korea

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

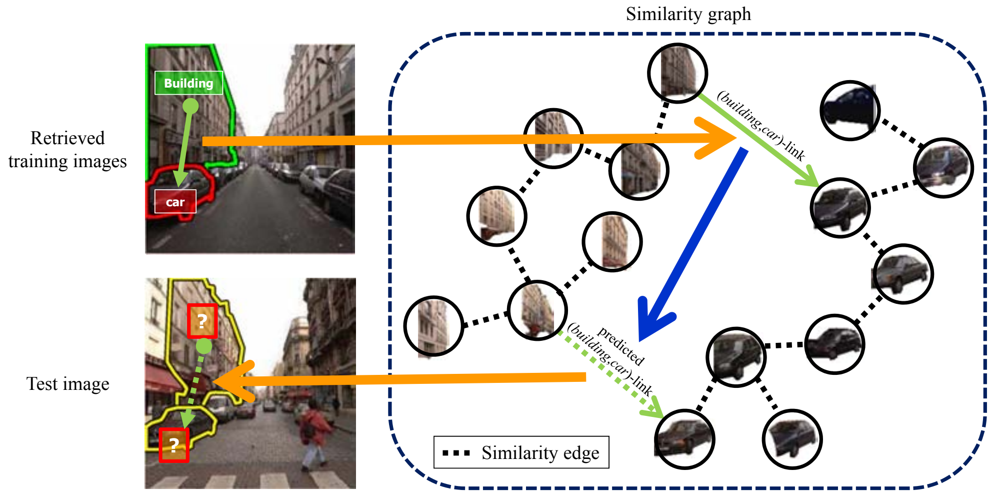


Figure 2. Illustration of our approach. The contextual relationship between the pair of the annotated regions is represented as (*building*,*car*)-link between the two corresponding nodes on the similarity graph. No link is constructed between the two regions from the test image because they are unlabeled. By applying link analysis techniques [18], our system predicts the strength of (*building*,*car*)-link between them based on node similarity.

to the visual appearance of objects to correctly label such unknown regions.

In this work, we present a novel approach for properly capturing the contextual relationships between two regions by considering the content of an input image. One difficulty is that learning such relations between all pairs of regions across whole object classes is computationally challenging. To overcome this problem, we propose a novel nonparametric exemplar-based context model. This nonparametric context model consists of a bunch of *context exemplars* which are basically annotated region pairs extracted from similar training images to the input image. From these context exemplars, we provides the novel interpretation of contextual relationships in a context link view and relates the problem of learning contextual relationships to the link prediction problem on a similarity graph of regions. The configuration of the similarity graph naturally reflects visual similarity between regions. On this similarity graph, all context exemplars can be compactly encoded in the form of context links. Moreover the similarity graph is usually sparse, so computation of learning contextual relationships can be done very efficiently.

The key contributions of this paper are as follows. (1) We establishes a novel context link view of contextual knowledge. (2) In this view, we formulate the problem of learning object relationships as graph-based link prediction problem which can be efficiently solved via state-of-the-art link analysis techniques [12, 18]. (3) Our system is nonparametric and exemplar-based, and therefore does not need to see whole training images to build a context model. Hence, it easily scales to large datasets with the tremendous number of images and object classes. Our system can also infer contextual relationships even from a single training image.

The rest of the paper is organized as follows. In Section 2, we review some relevant works. Section 3 presents our context model. Section 4 describes a region labeling algorithm using our context model. Section 5 provides experimental results and related discussion, followed by a conclusion in Section 6.

2. Related Work

There are two different types of approaches to object relationships. The first type focuses on the neighborhood interactions that captures the relation of two classes between nearby pixels/regions. To obtain it, various approaches have been proposed such as simple continuity preference [17], training classifier over pairwise features [7], and penalty term using co-occurrence statistics [26, 27]. However, the adjacent interactions is limited to modeling local properties of the image. Nevertheless, many existing nonparametric scene parsing methods [17, 26, 27] have employed neighboring relationships due to the scalability. The second type, on the other hand, models high-level relationships among objects by considering both long range and neighboring dependencies. This context model is typically represented by co-occurrence statistics or spatial relationships between object classes. Ravinovich *et al.* [22] incorporated co-occurrence statistics into the fully connected Conditional Random Field (CRF). Galleguillos *et al.* [6] proposed exploiting the information of relative location such as *above, beside,* or *enclosed* between object classes. Gould *et al.* [8] designed a more complex and informative relative location prior among object classes. Parikh *et al.* [21] differently learned co-occurrence statistics according to location and scale information. However, all these existing global context models rely on pixels/regions label prediction and are unable to incorporate visual appearance information effectively during context learning stage.

Jain *et al.* [11] proposed adaptively predicting "what" object relationships to consider and "how" to evaluate these relationships based on local and global image features. They learnt class-specific pairwise feature weights in a nonparametric manner, but they only consider simple relative position, overlap, and brightness. Different to Jain *et al.* [11], our approach relies on context link, allowing us to model complex object relationships directly associating to object classes.

Perhaps one of the most similar works to our approach is that of Malisiewicz and Efros [19]. They developed the Visual Memex graph with similarity and contextual edges. In contrast to [19], we builds the memex at query time only using matched images on global similarity level. Furthermore, our system reasons the strength of contextual relationships between regions, while [19] only predicts the category of a hidden object with some provided objects. This paves new promising way of representing and embedding higher-level semantic contextual relationships among objects in scene parsing and understanding.

3. Our Approach

3.1. Overview

For a query image, we first retrieve its best matched similar scenes in a large dataset using global descriptors analogous to several nonparametric scene parsing methods [17, 24, 26]. All pairs of the annotated regions in the retrieved scenes can be defined and exploited as *context exemplars*. A context exemplar is composed of a pair of regions and a pair of the corresponding object classes. It represents that a region with its particular object class supports the paired region to have its corresponding object class. For example, in Figure 2, the contextual relation from the region labeled as *building* to the region labeled as *car* forms a context exemplar. This means, when the former region is labeled as *building*, the latter region would be labeled as *car*. Note that this context exemplar can capture the global interaction between regions and is not limited to the local adjacent interaction.

Our goal is to estimate how much each region pair of the query image is consistent with the context exemplars from the retrieved images. For this, we first construct the similarity graph in which unlabeled regions from the test image and the annotated regions from the matched scenes are regarded as nodes. Each context exemplar is then encoded as a link between two nodes with the corresponding object classes on the similarity graph as illustrated in Figure 2. By applying the label propagation technique, a kind of semi-supervised learning method, the links between all nodes of the query image are constructed with their associated scores. Note that this label propagation method was originally proposed to solve the node classification problem [30]. After that, many researchers [12, 18] extended it to predict the relations among the nodes. In this work, we follow the approach of [18] because it is efficient compared to other methods [12]. Finally, the learned context scores are incorporated into the MRF framework for final labeling.

3.2. Retrieval System

A confident image set for the input test image is first extracted from a large training dataset because it is not scalable to consider all context exemplars from whole labeled training images. What we expect to have in the retrieval set are similar objects with consistent spatial arrangement compared to the test image. Hence, retrieval is done not only for computational efficiency but also for more informative region-based context learning.

Four different types of global image features are used: color histogram, spatial pyramid [16], gist [20], and tiny image [28]. For each feature, top-scored $T/4$ images according to the ranking scores are collected and used as the retrieval set similar to [26]. Having the best matches from each of the global features allows us to take into account various examples of scene context with the different views. All pairs of annotated regions in this retrieved set will form the context exemplars and serve as the source of region-level context learning.

3.3. Graph Construction

The k-nearest neighbor *similarity graph* is constructed between regions from both the test image I and the corresponding retrieved image set $\hat{\mathcal{I}}$. Each image is segmented into a number of regions based on the fast graph-based segmentation algorithm [4], and then each region is described by its appearance using selective shape, location, texture, color, and appearance features same as in [26]. The similarity graph is defined as a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ is a set of vertices that contains a set of regions $S_U = \{s_1, ..., s_M\}$ from the test image I and a set of regions $S_L = \{s_{M+1}, ..., s_N\}$ from the retrieved images $\hat{\mathcal{I}}$. Each vertex is connected to its k-nearest neighbor. A weight $w_{ij} \in \mathcal{W}$ is assigned to an edge $e_{ij} \in \mathcal{E}$, and is defined by the following similarity measure comparing two regions s_i and s_j based on Gaussian kernel:

$$w_{ij} = \prod_{H_k \in \mathcal{H}} \exp\left(-\frac{\|H_k(s_i) - H_k(s_j)\|}{\sigma_{H_k}}\right), \qquad (1)$$

where $H_k(s_i)$ is the feature vector of the k-th type for s_i, $\mathcal{H}$ represents the set of features, σ_{H_k} denotes the standard deviation of H_k, and all features are equally weighted.

3.4. Context Exemplar Description

In this step, the contextual relationships within the retrieved scenes are extracted in the form of contextual exemplars. Instead of counting co-occurrence or voting spatial arrangement between object classes, we simply extract all pairs of the annotated regions from the retrieved scenes and represent each pair as a context exemplar with the corresponding pair of object classes. More formally, given a set of classes $\mathcal{C} = \{c_1, c_2, ..., c_K\}$ (e.g. *sky, building, ..*) containing all existing classes in the corresponding retrieved image set $\hat{\mathcal{I}}$, the set of context exemplars for each class pair (c_a, c_b) is represented as

$$M^{ab} = \{(s_i, s_j) : G(s_i) = c_a, G(s_j) = c_b, s_i, s_j \in \hat{\mathcal{I}}_l\}, \qquad (2)$$

where $s_i, s_j \in \hat{\mathcal{I}}_l$ represents two regions s_i and s_j in the same image $\hat{\mathcal{I}}_l$ included in the retrieved image set $\hat{\mathcal{I}}$ and $G(s_i)$ represents the ground truth class of region s_i. Note that the order of all pairs (s_i, s_j) should be preserved since each context exemplar is assumed to have direction. Hence, based on region pair (s_i, s_j) labeled as (c_a, c_b), two context exemplars $(s_i, s_j) \in M^{ab}$ and $(s_j, s_i) \in M^{ba}$ are constructed. We hold $\mathcal{M} = \{M^{11}, M^{12}, M^{13}, ..., M^{KK}\}$ for all object class pairs and this contains the whole contextual relationships within the retrieved image set $\hat{\mathcal{I}}$ without loss of information.

Our key observation is that a context exemplar $(s_i, s_j) \in M^{ab}$ can be viewed as a directional (c_a, c_b)-type link between two nodes s_i and s_j on the similarity graph. We will refer to this link as the (c_a, c_b)-link. To transform all context exemplars into context link form, let $\mathcal{F}$ denote the set of $N \times N$ matrices with nonnegative entries. A matrix $\mathbf{F}^{ab} \in \mathcal{F}$ associates to (c_a, c_b)-links and $[\mathbf{F}^{ab}]_{ij}$ represents the strength of (c_a, c_b)-link between two nodes s_i and s_j. The strength close to 1 means high confidence of the existence of a link. On the other hand, the strength close to 0 means the absence of a link. We define $\mathbf{Q}^{ab} \in \mathcal{F}$ to represent the observed (c_a, c_b)-links within the retrieved images such that

$$[\mathbf{Q}^{ab}]_{ij} = \begin{cases} 1 & \text{if } (s_i, s_j) \in M^{ab} \\ 0 & \text{otherwise} \end{cases} . \qquad (3)$$

Now we have a set of context link $\mathcal{Q} = \{\mathbf{Q}^{11}, \mathbf{Q}^{12}, \mathbf{Q}^{13}, ..., \mathbf{Q}^{KK}\}$.

3.5. Context Link Prediction

Link prediction problem is a task of predicting how likely a link exists in a network. In this work, we consider a problem of predicting (c_a, c_b)-link among the pairs of nodes of S_U based on Q^{ab} consistent to the configuration of the similarity graph. For this, we adopt semi-supervised link propagation approach using node similarity similar to [12].

Procedure 1 Proposed Context Learning Algorithm

```
Input: Query image $I$
Output: Learned context scores $L(s_i, c_i, s_j, c_j)$
1: Retrieve the scene-level similar image set $\hat{\mathcal{I}}$
2: Generate superpixels $S_U$ of the query image $I$
3: Construct the similarity graph $\mathbf{W}$ of regions $S_U$ from $I$ and $S_L$ from $\hat{\mathcal{I}}$
4: Derive the matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ in which $\mathbf{D}$ is a diagonal matrix with its $(i,i)$-element equal to the sum of the $i$-th row of $\mathbf{W}$.
5: Extract the context exemplars $\mathcal{M}$ from $\hat{\mathcal{I}}$
6: Build the context link $\mathcal{Q}$
7: for each object class pair $(c_a, c_b)$ do
8:     Initialize $\mathbf{F}_c(1) = \mathbf{F}_r(1) = \mathbf{0}$
9:     ( Column-wise link propagation )
10:    Iterate $\mathbf{F}_c(t+1) = (1-c)\mathbf{L}\mathbf{F}_c(t) + c\mathbf{Q}^{ab}$ until convergence
11:    ( Row-wise link propagation )
12:    Iterate $\mathbf{F}_r(t+1) = (1-c)\mathbf{F}_r(t)\mathbf{L} + c\hat{\mathbf{F}}_c$ until convergence where $\hat{\mathbf{F}}_c$ indicates the limit of $\{\mathbf{F}_c(t)\}$
13:    Assign $L(s_i, c_i = c_a, s_j, c_j = c_b) = [\hat{\mathbf{F}}_r]_{ij}$ where $1 \le i, j \le M$ and $\hat{\mathbf{F}}_r$ denotes the limit of $\{\mathbf{F}_r(t)\}$
14: end for
```

We directly propagate (c_a, c_b)-links in Q^{ab} to the pairs of nodes of S_U and estimate the strength of them. We assume that all $\mathbf{Q}^{ab}$ is uncorrelated to each other, therefore, context link prediction problem can be solved by K^2 independent link propagation problems. We drop the ab suffix for clarity.

However, directly applying the approach of [12] to our context link prediction problem is impractical because it requires $O(N^4)$ times for a link propagation. Thus, we follow the strategy of the constraint propagation for spectral clustering [18]. We decompose the link propagation problem into two independent label propagation subproblems. First, the j-th column $\mathbf{Q}_{.j}$ serves as an initial configuration of two-class label propagation problem with respect to s_j. We will refer this process as a column-wise link propagation. The work of Zhou *et al.* [30] is employed to solve the label propagation problem with respect to s_j. All columns of $\mathbf{Q}$ are handled separately and the converged configuration $\hat{\mathbf{F}}_c$ (Step 10) is obtained. In practice, we observed that the columns of $\mathbf{Q}_{.j}$ within a retrieved image are exactly same. Therefore, only T, the number of retrieved images, of column-wise link propagation is required not N $(T \ll N)$.

Next, the i-th row of $[\hat{\mathbf{F}}_c]_{i.}$ is set as an initial configuration of two-class label propagation problem with respect to s_i. This is a row-wise propagation which works similar to the column-wise propagation. Practically, only what we want to obtain is the link information within the query image. Hence, row-wise link propagation with $(M < i \le N)$ is not necessary. After convergence of the row-wise iter-

Table 1. Performance comparison of our algorithm on Jain *et al.* [11] dataset and SIFT Flow dataset [17]. Per-pixel rates and average per-class rates in parentheses are presented.

	Jain *et al.* [11] Dataset	SIFT Flow dataset [17]
Jain *et al.* [11]	59.0 (-) [11]	-
Chen *et al.* [3]	75.6 (45) [3]	-
Liu *et al.* [17]	-	74.75 (-) [17]
Tighe and Lazebnik [26]	-	76.82 (29.38) [26]
Baseline classifier	77.62 (49.45)	73.35 (29.04)
Baseline MRF	76.48 (47.13)	74.08 (26.87)
Our (without ψ_i)	76.35 (45.72)	71.51 (30.84)
Our (with ψ_i)	**80.14 (53.25)**	**77.14 (32.29)**

ation (Step 12), the strength of (c_a, c_b)-link between two nodes s_i and s_j within the query image I is obtained.

Learning is independently performed for each Q^{ab} and repeated K^2 times. Each context learning is solved $O(kN^2)$ times on the k-nearest neighbor similarity graph ($k \ll N$) [18]. Therefore, the overall complexity of learning the context scores using our approach is $O(K^2N^2)$.

4. Inference

To assign labels to a set of regions S_U, the learned context scores $L(s_i, c_i, s_j, c_j)$ are incorporated to the fully connected MRF model. The fully connected model is proved to be effective for encoding the object interactions [6, 21, 22]. Similar to that of [6, 21, 22], we define the energy function of object class labels $\mathcal{C} = \{c_1, c_2, .., c_K\}$ as:

$$\mathcal{J}(\mathbf{c}) = \sum_{i=1}^{M} \psi_i(c_i) + \lambda \sum_{i,j=1}^{M} \phi_{ij}(c_i, c_j), \qquad (4)$$

where M is the number of regions in the test image I. The data term $\psi_i(c_i)$ represents the negative logarithm of the probability of class c_i given the region s_i. To obtain $\psi_i(c_i)$, we train discriminative classifiers from training dataset using visual features [26]. The smoothness term $\phi_{ij}(c_i, c_j)$ indicates pairwise contextual cost between the regions learned by our approach. This can be written as

$$\phi_{ij}(c_i, c_j) = -\log(\frac{1}{Z} L(s_i, c_i, s_j, c_j)), \qquad (5)$$

where $Z = \sum_{i=1}^{M} \sum_{c_i}^{K} L(s_i, c_i, s_j, c_j)$ is the normalization constant. Notice that the energy function is controlled by λ, which is the influence of the learned context scores. To minimize the MRF energy function, we applied α-expansion algorithm [2, 13] using the Quadratic Pseudo-Boolean Optimization (QPBO) algorithm [1, 23] which is publicly available[1].

5. Experiements

In this section, we report experimental results on two challenging datasets: the dataset of Jain *et al.* [11] and SIFT Flow dataset [17]. We evaluate the performance of the learned context scores and compare the accuracy of our approach both to a baseline and to recent state-of-the-art results. In each experiments, we evaluate four different models: a baseline classifier without MRF model; a baseline MRF with convential co-occurrence prior; our approach without unary potential. Our implementation is in MATLAB based on the available SuperParsing code[2]. We fix the parameters of our system with $T = 16, k = 10, c = 0.9, \lambda = 1$ in all experiments.

Baseline MRF: We evaluate the performance of the proposed approach against a conventional co-occurrence based model for object interaction. Following the most successful approaches [22], we incorporate the object class co-occurrence as local interaction into the fully connected MRF model. Hence, we design a baseline MRF model that has different form of the smoothness term to our model as

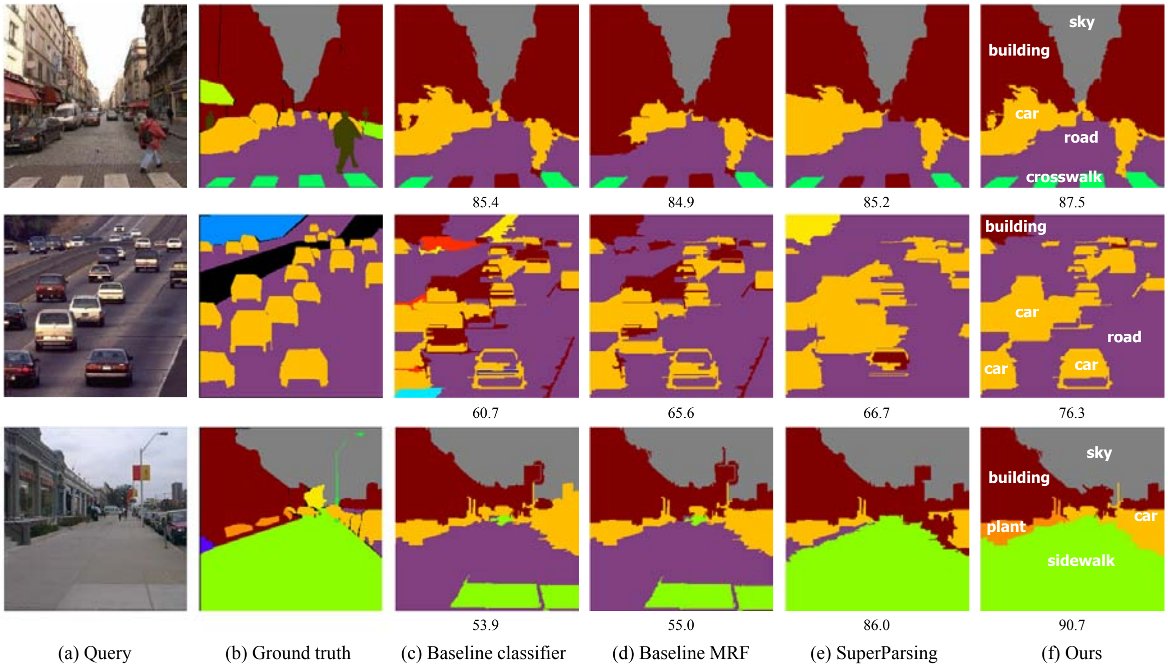
$$\phi_{ij}(c_i, c_j) = -\log(\frac{P(c_i|c_j) + P(c_j|c_i)}{2}) \times \delta[c_i \neq c_j], \qquad (6)$$

where $P(c_i|c_j)$ is the empirical probability of classes c_i and c_j co-occurring in the training images.

Jain *et al.* [11] Dataset: Jain *et al.* [11] dataset contains total 350 images randomly selected from LabelMe [25] dataset with 19 classes (250 training and 100 test images). We train boosted decision tree classifiers [10] for computing ψ_i terms. Per-pixel and per-class rates are presented in Table 1. Our system has an overall pixel-wise accuracy of 80.14% and a class-wise accuracy of 53.25%. We achieves pixel-wise 5% and class-wise 8% improvement over state-of-the-art performance [3]. Compared to the baseline MRF, our approach improves overall per-pixel rates by about 4% and this result clearly shows the advantage of our approach.

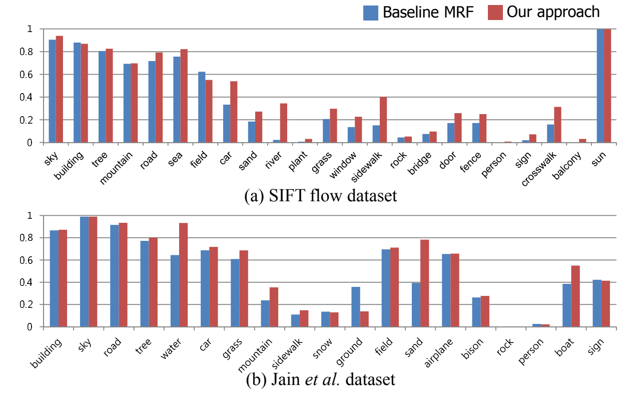
[1] http://pub.ist.ac.at/~vnk/software.html

[2] http://www.cs.unc.edu/~jtighe/Papers/ECCV10/index.html



(a) Query (b) Ground truth (c) Baseline classifier (d) Baseline MRF (e) SuperParsing (f) Ours

Figure 3. Representative results from the SIFT Flow dataset. Column (a) shows the query image to be labeled and Column (b) represents the ground truth of (a). Column (c), (d), (e), and (f) show the prediction of the baseline classifier, baseline MRF models, SuperParsing [26], and our approach with unary potential, respectively. The numbers under each image indicate pixel-wise accuracy (%) on that image. Crosswalk is appeared in the first row, building is removed without smoothing in the second row, and sidewalk and plant are recovered in the last row. Obviously, implausible baseline classifier results are appropriately corrected based on the learned context scores. These figures are best viewed in color.



(a) SIFT flow dataset

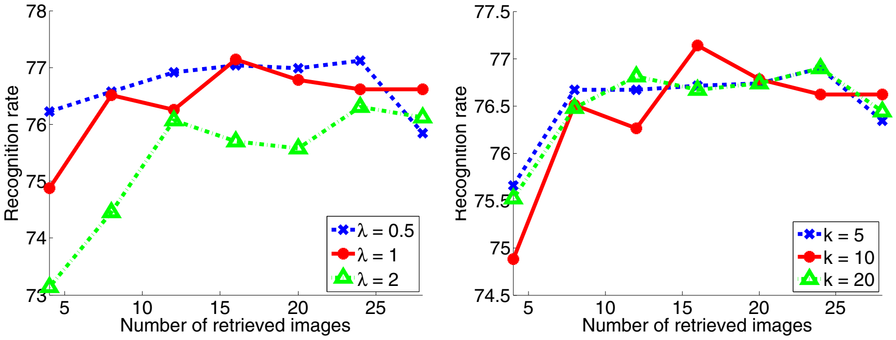
(b) Jain *et al.* dataset

Figure 4. The per-class recognition rate of our system compared with baseline MRF on (a) SIFT flow dataset and (b) Jain *et al.* [11] dataset.

More importantly, baseline MRF drops per-class rates since the conventional context models smooth away smaller object classes. On the other hands, our approach does not suffer from such problem and even improves per-class rates by 3%.

SIFT Flow dataset: The SIFT Flow dataset provided by Liu *et al.* [17] consists of 2,688 images of outdoor scenes. The dataset provides ground truth labels hand-annotated by LabelMe users. Liu *et al.* [17] split this dataset into 2,488 training images and 200 test images, and selected top 33 object categories as semantic labels. For comparison, the same training/test split is used as [17, 26]. To obtain ψ_i terms, we employ nonparametric nearest-neighbor classifiers [26, 27]. Our system achieves an overall pixel-level accuracy of 77.49% and a per-class accuracy of 32.29%. Figure 4 (a) shows that our per-class rate on the SIFT Flow dataset is significantly better than that of the baseline MRF.

Next, we validate our system by varying the parameters including the number of retrieved images T, the feature combination, k of k-nearest neighbor, and the influence of context scores λ. First, we fix $k = 10$, use all features, and plot the recognition rate as a function of T in Figure 5 (a) with different λ. The recognition rate increases as more retrieved images are used. However, the recognition rate slightly drops continue to add retrieved images. Additionally, it is observed that strongly enforcing contextual consistency increases ambiguities and degenerates the performance. The maximal performance is achieved when $T = 16$ and $\lambda = 1$. Second, we fix $\lambda = 1$, use all features, and plot the recognition rate as a function of T in Figure 5 (b) with different k. Clearly, appropriate number of re-



(a) (b)

Region feature	Rate (%)
SIFT	75.27 (28.29)
+ Texture	75.62 (29.28)
+ Location	76.49 (30.65)
+ Shape	76.82 (30.68)
+ Appearance	76.80 (30.97)
+ Color	77.14 (32.29)

(c)

Figure 5. (a): Recognition rate as a function of the number of the retrieved images T and the influence of our model λ. (b): Recognition rate as a function of the number of the retrieved images T and the k of the visual similarity graph. (c): Feature evaluation on the SIFT Flow dataset.

Table 2. Average computation time in second.

	Jain *et al.* Dataset	SIFT Flow Dataset
Image size	640 × 480 (few exception)	256 × 256
Average N (the number of regions)	4243	1005
Average K (the number of object class)	15	11
	Time (second)	
Graph Construction	23.92	4.22
Context Link Prediction	155.51	18.09
Inference	8.83	0.79

trieved images is needed to achieve accurate context consistency. The maximal performance is achieved when $T = 16$ and $k = 10$. Finally, Figure 5 (c) shows recognition rates with region features added consecutively. Notice that it is arranged in order of increasing per-class rate and the SIFT histogram is the strongest feature in our system similar to the result of [26].

The computation time of our algorithm is shown in Table 2. All experiments were run on a standard PC with 3.0 GHz Intel quadcore CPU and 8 GB RAM. For both datasets, we fixed our parameters to $T = 16, k = 10, \lambda = 1$. It means that total $T + 1 = 17$ images are used to construct a similarity graph. Since our algorithm requires $O(K^2N^2)$ times, increasing K and N makes our algorithm significantly slow.

6. Conclusion

We have presented a nonparametric exemplar-based context model in which object relationships are explicitly captured. A graph-based context representation is proposed to efficiently transfer contextual relationships from training images to a query image. This allows jointly modeling visual appearance and context. Our novel approach helps to overcome the limitation of conventional context models relying on object label agreement and gives richer appearance-based context information. Moreover, the learned object relationships can be incorporated into any region-based scene labeling approaches as an additional cue. One of the main limitations of our model is that it considers all relations between regions as equally important. Clearly, there might be implausible or unimportant context exemplars, but our model cannot eliminate them. Our future work is to overcome this problem and extend our system to the multiple segmentation framework.

References

[1] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123:155–225, 2002. 5

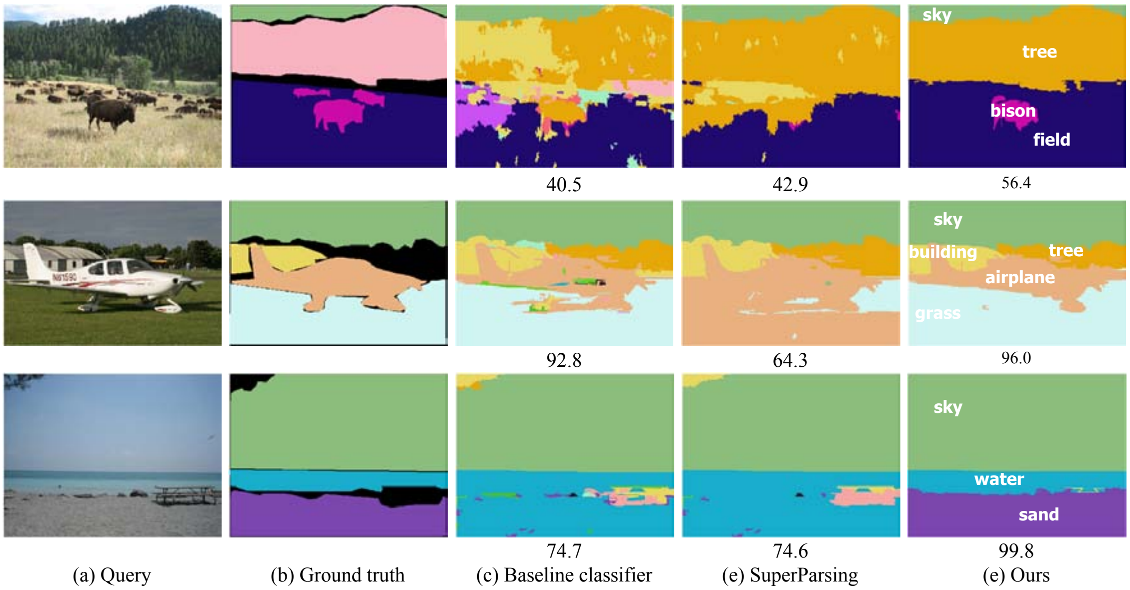
[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26:1124–1137, September 2004. 5

[3] X. Chen, A. Jain, A. Gupta, and L. S. Davis. Piecing together the segmentation jigsaw using context. In *Computer Vision and Pattern Recognition(CVPR)*, 2011. 5

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59:167–181, 2004. 3

[5] C. Galleguillos, B. McFee, S. Belongie, and G. R. G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2010. 1

[6] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 2, 5



(a) Query (b) Ground truth (c) Baseline classifier (e) SuperParsing (e) Ours

Figure 6. Example results from Jain *et al*. dataset. Column (a) shows the query image to be labeled and Column (b) represents the ground truth of (a). Column (c), (d), (e) shows the prediction of the baseline classifier, SuperParsing [26], and our approach with unary potential, respectively. The numbers under each image indicate pixel-wise accuracy on that image.

[7] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *International Conference on Computer Vision(ICCV)*, 2009. 1, 2

[8] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *Int. J. Comput. Vision*, 80(3):300–316, 2008. 1, 2

[9] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *European Conference on Computer Vision (ECCV)*, 2008. 1

[10] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision*, 75:151–172, 2007. 1, 5

[11] A. Jain, A. Gupta, and L. S. Davis. Learning what and how of contextual models for scene labeling. In *European Conference on Computer Vision (ECCV)*, 2010. 1, 3, 5, 6

[12] H. Kashima, T. Katoy, Y. Yamanishiz, and M. Sugiyama. Link propagation: A fast semi-supervised learning algorithm for link prediction. In *SIAM International Conference on Data Mining*, 2009. 2, 3, 4

[13] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26:147–159, 2004. 5

[14] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *International Conference on Computer Vision(ICCV)*, 2005. 1

[15] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision (ECCV)*, 2010. 1

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR)*, 2006. 3

[17] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 3, 5, 6

[18] Z. Lu and H. H. Ip. Constrained spectral clustering via exhaustive and efficient constraint propagation. In *European Conference on Computer Vision (ECCV)*, 2010. 2, 3, 4, 5

[19] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *Advances in Neural Information Processing Systems*, December 2009. 3

[20] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*, 2006. 3

[21] D. Parikh, C. L. Zitnick, and T. Chen. From appearance to context-based recognition:dense labeling in small images. In *Computer Vision and Pattern Recognition(CVPR)*, 2008. 1, 2, 5

[22] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision(ICCV)*, 2007. 1, 2, 5

[23] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary mrfs via extended roof duality. In *Computer Vision and Pattern Recognition(CVPR)*, 2007. 5

[24] B. C. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *Advances in Neural Information Processing Systems*, 2007. 3

[25] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77:157–173, May 2008. 5

[26] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision(ECCV)*, 2010. 2, 3, 5, 6, 7, 8

[27] J. Tighe and S. Lazebnik. Understanding scenes on many levels. In *International Conference on Computer Vision(ICCV)*, 2011. 2, 6

[28] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30:1958–1970, 2008. 3

[29] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, pages 1401–1408. MIT Press, Cambridge, MA, 2005. 1

[30] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2004. 3, 4