A Confidence Ranked Co-Occurrence Approach for Accurate Object Recognition in Highly Complex Scenes

Pelin Angin, Bharat Bhargava
Department of Computer Science, Purdue University, USA
{pangin, bb}@cs.purdue.edu

Abstract

Real-time and accurate classification of objects in highly complex scenes is an important problem for the Computer Vision community due to its many application areas. While boosting methods with the sliding window approach provide fast processing and accurate results for particular object categories, they cannot achieve the desired performance for more involved categories of objects. Recent research in Computer Vision has shown that exploiting object context through relational dependencies between object categories leads to improved accuracy in object recognition. While efforts in collective classification in images have resulted in complex algorithms suitable for offline processing, the real-time nature of the problem requires the use of simpler algorithms. In this paper, we propose a simple iterative algorithm for collective classification of all objects in an image, exploiting the global co-occurrence frequencies of object categories. The proposed algorithm uses multiple detectors trained using Gentle Boosting, where the category of the most confident estimate is propagated through the co-occurrence relations to determine the categories of the remaining unclassified objects. Experiments on a real-world dataset demonstrate the superiority of our approach over using Gentle Boosting alone as well as classic collective classification approaches modeling the full joint distribution for each object in the scene.

Keywords: Computer vision, Object recognition, Co-occurrence, Confidence, Real-Time.

1 Introduction

Accurate object categorization is an important problem, given the increasing demand for machine vision components in many systems today. Fields as diverse as assistive technologies, security systems at the airports and part tracking systems at factories, among others need automatic object categorization to obviate the need for human processing of the available information to achieve particular tasks they are designed for. Although decades of work by the Computer Vision community has contributed to the development of robust object classification algorithms, which work well in relatively well-defined settings with a few objects, the problem of accurately classifying all objects in a highly complex scene in real-time remains largely unsolved.

Research in object categorization has shown that boosted detectors can be effective in detecting particular classes of objects such as human faces [14] and cars [11], while other classes of objects having higher variance in shape, color and texture have proven much more difficult to detect. Training these detectors with few samples results in a high miss rate for the category they are trained for, which could provide important clues about the setting, whereas using training samples from a more diverse collection usually causes a high rate of false alarms. However, boosted detectors still prevail in object recognition due to their fast image processing capability, which is required by real-time applications.

Studies in statistical relational learning have focused on the importance of mining graphs modeling relationships between different entities to extract useful information, which allows for accurate classification of nodes with unknown class labels. Recent studies in object recognition have also focused on exploiting the context of an object in addition to the intrinsic features such as color, texture, shape cues and edges. Context has been defined and used in a few different ways by these studies, including determination of the location of an object in a scene as in [7]; enforcement of rules on the relative positions of things in a scene as in [9]; exploiting texture regions to help detection of objects as in [6] and using the co-occurrence frequency of multiple object categories in the same scene, which is the definition of context we will be using in this paper.

In this paper, we propose an algorithm for fast and accurate recognition of all objects in a highly complex scene with no prior knowledge of any object's category. The proposed algorithm exploits the efficiency of Gentle Boosting detectors [3] trained on multiple object categories and the power of relational autocorrelation to determine the possibility of co-occurrence of objects of these categories in the same scene. The main difference of the proposed approach from classic iterative collective classification algorithms is the use of class-label fixing for the object with the most confident class probability estimate at each iteration, which is propagated to neighboring nodes through global level contingency coefficients (i.e., co-occurrence frequencies) to refine their probability estimates. By

*Corresponding author: Pelin Angin; E-mail: pangin@cs.purdue.edu
DOI: 10.6138/JIT.2013.14.1.02

employing a confidence-ranked approach to fix class labels, the proposed approach provides improved classification accuracy as the most confident estimates are not updated further and the class labels of the neighboring nodes, which either go undetected or undetermined (due to low confidence levels for each category) by the boosting algorithm, are decided based on the co-occurrence rules extracted from real world data.

The rest of the paper is organized as follows: Section 2 includes a brief summary of related work in object recognition and use of object context for this task; Section 3 discusses the ubiquity of relational autocorrelation and its use in within-network classification; Section 4 describes the proposed object recognition model; Section 5 provides experiment results on the LabelMe image dataset and Section 6 concludes with future work directions.

2 Related Work

Recent research in the Computer Vision community has focused on the use of contextual clues in addition to the intrinsic features of objects for accurate object classification. Semantic relations including information about the interactions of objects in a scene were used by several classification models [2][5][8]. Among others, classification models to capture the local context of objects by considering the objects surrounding the target object were developed and shown to increase the accuracy of class predictions [10][13][15]. Torralba et al. [12] introduced Boosted Random Fields to model the correlation between different objects in a scene by exploiting the power of boosting and conditional random fields for accurate classification. Galleguillos et al. introduced a classification model [4] using a conditional random field (CRF) to maximize object label agreement according to both semantic and spatial relevance, where they model relative location between objects using simple pairwise features.

The models listed above mainly involve complex algorithms to achieve accurate classification, hence are suitable for offline processing of large image collections. Especially the models considering involved relations such as those based on rules about pair-wise location relations involve complex computation. The problem we are addressing in this paper is the real-time recognition of objects in a complex scene. Thus, our purpose here is to keep the model as efficient as possible to meet the real-time requirement. Close to our work is the work of Yun et al. [16], which models the relations between different objects in a graphical model for real-time recognition. However, unlike our approach, their approach is based on a time transition model, requiring follow-up of the scene for a period of time.

3 Relational Autocorrelation

Relational autocorrelation refers to a statistical dependency between values of the same variable on related objects. For a graph $G = (V; E)$, where each node $v \in V$ represents an object and each edge $e \in E$ represents a binary relation, autocorrelation is measured for a set of instance pairs P_R related through paths of length l in a set of edges E_R:

$$P_R = \{(v_i v_j): e_{ik1}, e_{k1k2}, \ldots, e_{klj} \in E_R\} \quad (1)$$

where $E_R = \{e_{ij}\} \subseteq E$. It is the correlation between the values of a variable X on the instance pairs $(v_i x, v_j x)$ such that $(v_i, v_j) \in P_R$ [1]. Autocorrelation is a nearly ubiquitous characteristic of relational datasets.

When there are dependencies among the class labels of related instances, relational models can exploit those dependencies by including related class labels as dependent variables in the model. Then the probability distribution for the target class label (Y) of an instance i can be conditioned not only on the attributes (X) of i in isolation, but also on the attributes and class labels of instances ($R = \{1, \ldots, r\}$) related to i:

$$P(y^i|x^i, \{x^1, \ldots, x^r\}, \{y^1, y^r\}) = P(y^i|x^i_1, \ldots, x^i_m, x^1_1, \ldots, x^1_m, \ldots, x^r_1, \ldots, x^r_m, y^1, \ldots, y^r) \quad (2)$$

Relational autocorrelation used frequently in classifying relational (graph) data is referred to as co-occurrence frequency in the Computer Vision community. For the task of classifying objects in a scene, the scene is assumed to be a graph comprised of objects, which are all connected to each other via edges.

4 Boosting and Co-Occurrence Based Object Recognition

4.1 Basic Classification Model

The basic classification model we use in this work is a local classification model employing a simple Naive Bayes classification scheme, where the probability of an object's belonging to a particular class is conditioned on the characteristics of the other objects related to that object (which we call the neighbor objects) and each neighbor is conditionally independent given the class.

The model defines the probability that an object i belongs to category y as:

$$p(y^i|N(i)) \alpha p(y) * \prod_{j \in N(i)} p(y|j), \quad (3)$$

$N(i)$: The set of labeled neighboring objects of i;
$p(y)$: The prior probability of class label y for object i;
$p(y|j)$: The conditional probability of y given that the object co-occurs with the object j. This conditional probability is looked up from the contingency table including the co-occurrence frequencies for each pair of object categories, pre-built based on the training data. A graphical model of the scene is constructed by allocating a node for each object in the scene, and placing them at the locations on a 2-dimensional plane corresponding to their coordinates in the original image.

All of the following classification algorithms are based on this model, however differ in the neighbor set used to determine the class label of a particular object. Gentle Boosting is used to train separate detectors for each object category and the detector output of the different detectors for each object is used to determine the prior probability for each class label for that object (if there is no detection of an object by a particular detector, the prior probability is set to a very low value).

4.2 Iterative Full Joint Collective Classification

Iterative collective classification is a popular approach for classifying entities linked through various relations. In the iterative collective classification algorithm implemented as basis for comparison in this paper, at each iteration of the algorithm, the probability estimates for each class label for each object is updated based on the probability estimates of all the remaining objects in the previous iteration of the algorithm. The algorithm runs until the probability estimates are converging or after a threshold number of iterations have been performed. The full joint estimate takes into consideration the probability of every possible class label for each neighbor of a specific object, requiring a complex computation. This scheme has the disadvantage that the algorithm could take a long time to converge, which is not suitable for real-time applications.

4.3 Confidence-Ranked Iterative Classification

The confidence-ranked iterative classification approach that we propose in this paper starts by setting the prior probabilities to the outputs of the detectors trained for each class (and normalizing as necessary) as in the iterative full joint collective classification algorithm described above. At each iteration of the algorithm, the most confident class probability estimate among those of all yet unclassified objects is found, and the class label of that object is set to the class with this highest probability. Then, the class probability estimates of the remaining unclassified objects are updated based on the co-occurrence frequencies of the classes with the class label of the last classified object. Pseudocode for the algorithm is given in Figure 1. The

```
/*Initialization*/
/*Here m is the number of possible class labels*/
U ← set of unlabeled objects in scene
for i ∈ U do
  for c = 1 → m do
    P(y^i_c) ← Gentle boosting probability of y_c for i
  end for
end for
labeledObject ← object with highest probability of a class
label in U
U ← U – labeledObject

/*loop until all objects are categorized*/
while U ≠ φ do
  for i ∈ N(labeledObject) do
    if i ∈ U then
      for c = 1 → m do
        P(y^i_c) ← P(y^i_c) * P(y^i_c|labeledObject)
      end for
      Normalize class probabilities for i
    end if
  end for
  labeledObject ← object with highest probability of a
  class label in U
  U ← U – labeledObject
end while
```

Figure 1 Pseudocode for Confidence-Ranked Iterative Classification Algorithm

algorithm runs until all objects are classified, which limits the number of iterations to the number of objects in the scene. The simplicity and speed of this algorithm makes it ideal for real-time applications.

4.4 Nearest-Neighbor Based Classification

A modified version of the confidence-ranked iterative classification algorithm described above propagates the fixed class label information only to a limited set of neighbors in the immediate vicinity of the classified object. The distance between two objects in this scheme is calculated as the Euclidean distance between the centers of the bounding boxes of the two objects. Angin et al. [1] have shown that exploiting local dependencies between the class labels of neighboring nodes leads to accurate results only when the information at the local level is sufficient. Experiments with different neighborhood sizes in section 5 demonstrate that algorithms restricted to the immediate locality of an object for co-occurrence information do not perform well in cases where the local information is not sufficient for accurate classification (e.g., when the immediate neighbors have low confidence values).

5 Experimental Evaluation

Real-world experiments were performed on a subset of the LabelMe image dataset (http://labelme.csail.mit.edu/), consisting of indoor scenes (office, living room, bathroom, kitchen etc). A subset of the images used is seen in Figure 2. The images for both the training and test sets were selected such that they contain as many objects as possible, allowing us to see the efficiency of the proposed algorithm on highly complex scenes. 100 images for each of 12 most prevalent object classes were used to train detectors using Gentle Boosting as well as forming the co-occurrence frequency table to be used by the different classification algorithms. Tests were performed on 2,067 images. For each test image, only objects of the 12 categories were considered as in the scene, the bounding boxes of which were extracted using the LabelMe toolbox.

The precision values for different classes of objects and the three classification models described above are given in Table 1 and the recall values are given in Table 2. As seen in the tables, while the iterative full joint collective classification method achieves high precision for most classes of objects, it has a big degrade in the recall values over using boosting alone. While our proposed approach achieves better accuracy than the approach using Gentle Boosting alone, it has a moderate decrease in recall for most classes and even allows for increased recall for some classes like the cupboard and the mouse. This is due to the use of confidence based fixing of class labels at each iteration, allowing us to keep the good estimates of the boosting algorithm, which are informative for the classification of the remaining objects. To see the overall classification effectiveness of the described classification models, we also calculated the F1-measure for each classifier for different object classes (given in Table 3). The F1-measure is a popular measure of the goodness of a classifier in the Information Retrieval community and is calculated as the harmonic mean of the precision and recall values for each category. The F1-measure also demonstrates the superiority of the proposed approach over the other two approaches compared to. We performed paired, two-tailed statistical t-tests to compare the F1-Measure for the proposed approach and the other two models. The p-values of 0:0018 and 0:0003 for the tests with the Boosting only and the full joint collective classification schemes respectively demonstrates the significance of the gainings by using the proposed approach.

We also performed experiments to see the effect of the neighborhood size to which the class label of the most confident estimate at each step is propagated. Figure 3 gives the precision values (on vertical axis) for each object category for different values of the neighborhood size k and

Figure 2 Sample Images from the Labelme Dataset

Table 1 Precision Values for Different Classification Models

Object category/ model	Boosting only	Full joint collective	Conf. ranked iterative
Chair	0.43	0.19	0.25
Lamp	0.33	1.00	0.45
Table	0.13	0.23	0.19
Monitor	0.33	0.97	0.47
Keyboard	0.20	1.00	0.40
Sink	0.19	0.95	0.36
Bed	0.32	1.00	0.52
Faucet	0.07	0.92	0.13
Cupboard	0.19	0.75	0.32
Mouse	0.12	0.89	0.30
Plant	0.18	0.88	0.31
Vase	0.04	0.00	0.05

Table 2 Recall Values for Different Classification Models

Object category/ model	Boosting only	Full joint collective	Conf. ranked iterative
Chair	0.25	0.98	0.58
Lamp	0.26	0.06	0.25
Table	0.08	0.01	0.08
Monitor	0.59	0.08	0.60
Keyboard	0.33	0.08	0.38
Sink	0.34	0.12	0.30
Bed	0.58	0.12	0.51
Faucet	0.17	0.05	0.13
Cupboard	0.16	0.00	0.40
Mouse	0.16	0.02	0.39
Plant	0.21	0.04	0.17
Vase	0.08	0.00	0.05

Table 3 F1-Measure for Different Classification Models

Object category/ model	Boosting only	Full joint collective	Conf. ranked iterative
Chair	0.31	0.31	0.35
Lamp	0.29	0.11	0.32
Table	0.10	0.02	0.12
Monitor	0.43	0.14	0.53
Keyboard	0.24	0.15	0.39
Sink	0.25	0.21	0.33
Bed	0.41	0.21	0.52
Faucet	0.10	0.09	0.13
Cupboard	0.17	0.01	0.36
Mouse	0.14	0.04	0.34
Plant	0.19	0.08	0.22
Vase	0.05	0.00	0.05

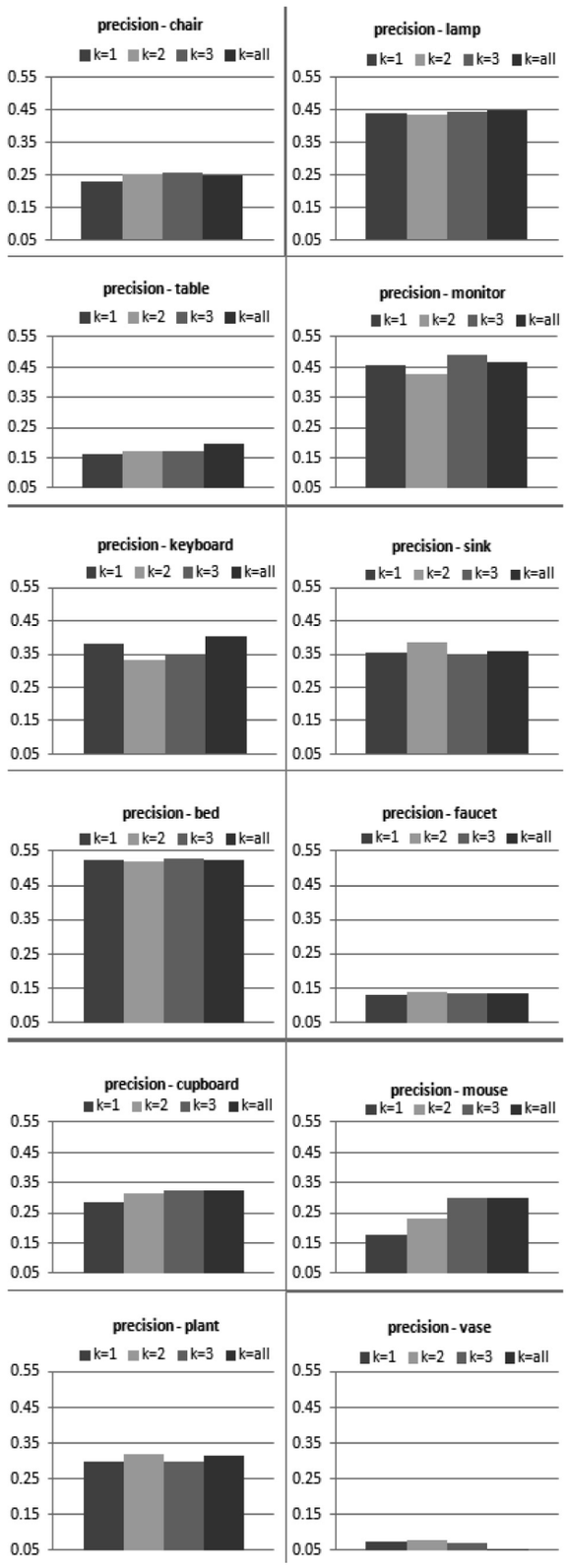

Figure 3 Precision Values for Different Neighborhood Sizes in the Confidence-Ranked Iterative Classification Algorithm

Figure 4 provides the recall values for these experiments. In the graphs, k = 1 stands for the experiments where the fixed class label of an object (the most confident estimate of an iteration of the algorithm) is only allowed to influence the class probability estimate of the nearest neighbor of that object. Likewise, k = 2 stands for influencing 2 nearest neighbors and k = all stands for the standard algorithm proposed, where the class label information is propagated to all objects in the scene.

As seen in the graphs, while the precision values for each neighborhood size are close to each other for most object categories, propagating label information only to the nearest neighbor causes significant degrades in recall performance for classes such as the cupboard, keyboard and mouse. We also observe that there is no significant difference between the performances of a neighborhood size of 3 and considering all objects in the image. These results imply that using a single neighbor does not allow for the propagation of the accurate classification results. This justifies our decision to use the global level of co-occurrences. Using the global co-occurrence frequencies instead of being restricted to local ones also helps capture the global scene description without apriori knowledge about it.

6 Conclusion

In this paper, we proposed a simple iterative algorithm for collective classification of all objects in an image in real-time, exploiting the global co-occurrence frequencies of object categories. The algorithm is based on using multiple detectors trained using Gentle Boosting, where the category of the most confident estimate is propagated through the co-occurrence relations to refine the probability estimates of the categories of the unclassified objects. Experiments on the LabelMe dataset of real-world images demonstrate the superiority of our approach over using Gentle Boosting alone as well as classic collective classification approaches modeling the full joint distribution for each object in the scene. We also demonstrated the results of using different neighborhood sizes in the proposed algorithm to see the local-global co-occurrence effects. Future work will involve incorporation of other types of context into the algorithm for increased accuracy and efforts to avoid degrade in recall rates over using boosting alone.

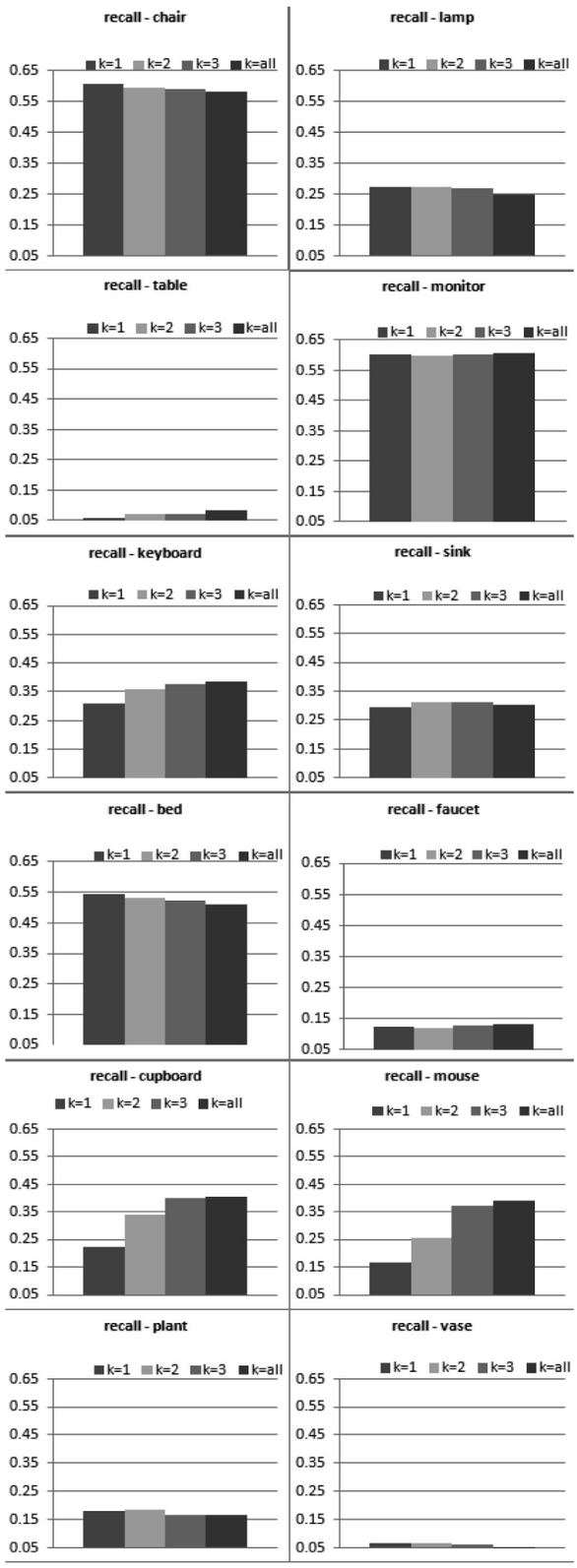

Figure 4 Recall Values for Different Neighborhood Sizes for the Confidence-Ranked Iterative Classification Algorithm

References

[1] Pelin Angin and Jennifer Neville, *A Shrinkage Approach for Modeling Non-stationary Relation Autocorrelation*, *Proc. IEEE ICDM*, Pisa, Italy, December, 2008, pp.707-712.

[2] Peter Carbonetto, Nando de Freitas and Kobus Barnard, *A Statistical Model for General Contextual Object Recognition*, *Proc. European Conference on Computer Vision (ECCV)*, Prague, Czech Republic, May, 2004, pp.350-362.

[3] Jerome Friedman, Trevor Hastie and Robert Tibshirani, *Additive Logistic Regression: A Statistical View of Boosting*, *Annals of Statistics*, Vol.28, No.2, 1998, pp.337-407.

[4] Carolina Galleguillos, Andrew Rabinovich and Serge Belongie, *Object Categorization Using Co-occurrence, Location and Appearance*, *Proc. IEEE CVPR*, Anchorage, AK, June, 2008, pp.1-8.

[5] Xuming He, Richard S. Zemel and Miguel A. Carreira-Perpinan, *Multiscale Conditional Random Fields for Image Labeling*, *Proc. IEEE CVPR*, Washington, DC, June, 2004, pp.695-702.

[6] Geremy Heitz and Daphne Koller, *Learning Spatial Context: Using Stuff to Find Things*, *Proc. European Conference on Computer Vision (ECCV)*, Marseille, France, October, 2008, pp.30-43.

[7] Kevin Murphy, Antonio Torralba and William T. Freeman, *Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes*, *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December, 2003, http://books.nips.cc/papers/files/nips16/NIPS2003_VM02.pdf

[8] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora and Serge Belongie, *Objects in Context*, *Proc. IEEE ICCV*, Rio de Janeiro, Brazil, October, 2007, pp.1-8.

[9] Amit Singhal, Jiebo Luo and Weiyu Zhu, *Probabilistic Spatial Context Models for Scene Content Understanding*, *Proc. IEEE CVPR*, Madison, WI, June, 2003, pp.235-241.

[10] Thomas M. Strat and Martin A. Fischler, *Context-Based Vision: Recognizing Objects Using Information from Both 2D and 3D Imagery*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.13, No.10, 1991, pp.1050-1065.

[11] Antonio Torralba, *Contextual Priming for Object Detection*, *International Journal of Computer Vision*, Vol.53, No.2, 2003, pp.169-191.

[12] Antonio Torralba, Kevin P. Murphy and William T. Freeman, *Contextual Models for Object Detection Using Boosted Random Fields*, *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December, 2004, pp.1401-1408.

[13] Jacob Verbeek and Bill Triggs, *Scene Segmentation with Crfs Learned from Partially Labeled Images*, *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December, 2007, pp.1553-1560.

[14] Paul Viola and Michael Jones, *Robust Real-Time Face Detection*, *International Journal of Computer Vision*, Vol.57, No.2, 2004, pp.137-154.

[15] Lior Wolf and Stanley Bileschi, *A Critical View of Context*, *International Journal of Computer Vision*, Vol.69, No.2, 2006, pp.251-261.

[16] Woo-Han Yun, Sung Yang Bang and Daijin Kim, *Real-Time Object Recognition Using Relational Dependency Based on Graphical Model*, *Pattern Recognition*, Vol.41, No.2, 2008, pp.742-753.

Biographies

Pelin Angin is a PhD student at the Department of Computer Science at Purdue University. She received her BS degree in Computer Engineering at Bilkent University, Turkey in 2007. Her research interests lie in the fields of Mobile-Cloud Computing, Cloud Computing Privacy and Data Mining.

Bharat Bhargava is a professor of the Department of Computer Science with a courtesy appointment in the School of Electrical & Computer Engineering at Purdue University. He is a fellow of IEEE and IETE. He serves on editorial boards of ten international journals and is editor-in-chief of three journals.