Door Detection via Signage Context-based Hierarchical Compositional Model

Cheng Chen and Yingli Tian
The City College, City University of New York
New York, NY 10031
{cchen1,ytian}@ccny.cuny.edu

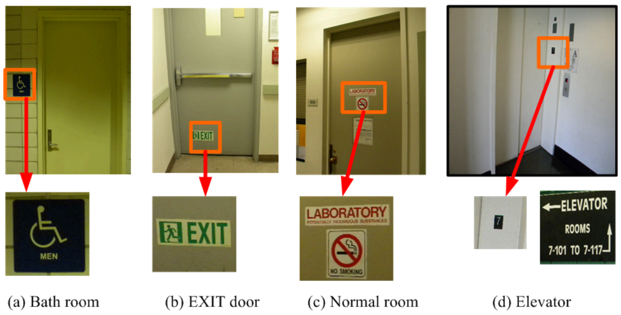
Abstract

Door detection by using wearable cameras helps people with severe vision impairment to independently access unknown environments. The goal of this paper is to robustly detect different doors and classify them as office doors, elevators, exits, etc. These tasks are challenging due to the factors: 1) small inter-class variations of different objects such as office doors and elevators, 2) only part of an object is captured due to occlusions or continuous camera moving of a mobile system. To overcome the above challenges, we propose a Hierarchical Compositional Model (HCM) approach which incorporates context information into the model decomposition process of a part-based HCM to handle partially captured objects as well as large intra-class variations in different environments. Our preliminary experimental results demonstrate promising performance on doors detection over a wide range of scales, view points, and occlusions.

1. Introduction

Independent travel is well known to present significant challenges for individuals with severe vision impairment, thereby reducing quality of life and compromising safety. Computer vision technology has the potential to assist blind individuals in independently accessing unfamiliar environments. There have been many efforts to study blind navigation and wayfinding with the ultimate goal of developing useful travel aids for blind people, but very few have met with more than limited success. The most useful and accepted independent travel aids remain the Hoover cane and the guide dog. While GPS-guided electronic wayfinding aids show much promise in outdoor environments, there is still a lack of orientation and navigation aids to help people with severe vision impairment to independently find doors, rooms, elevators, stairs, bathrooms, and other building amenities in unfamiliar indoor environments.

These important tasks are very challenging due to the factors: 1) small inter-class variations of different objects such as office doors and elevators, 2) only part of an object is captured due to occlusions or continuous camera moving of a mobile system.



(a) Bath room (b) EXIT door (c) Normal room (d) Elevator

Figure 1. Indoor objects (top row) and the associated contextual information (bottom row)

Considering the high appearance similarity of different doors such as office doors and elevator doors, it is very challenging to design object models which are flexible enough to cover large intra-class variation, while still retain good discriminative power between the object classes. Moreover, for mobile systems, it is unrealistic to expect all input images with intact object boundary. Therefore, reliable occlusion handling is also one of the key issues for a successful object detection/localization algorithm running on mobile systems.

In order to reliably detect doors, such as an elevator, an office, a bathroom and an exit (as shown in the first row of Figure 1), we propose a new part-based hierarchical compositional Model (HCM) approach by incorporating context information (the second row of Figure 1).

In this work, we use a dichotomy of useful information for object detection: appearance information and context information. Appearance information will be represented by an object model (e.g. part-based hierarchical compositional Model in this work). Context information includes distance between camera and objects, directional and locational signage as shown in the Figure 2. Distance between camera and objects is an important context cue for the possible partial-capture. The directional signage (the first row of

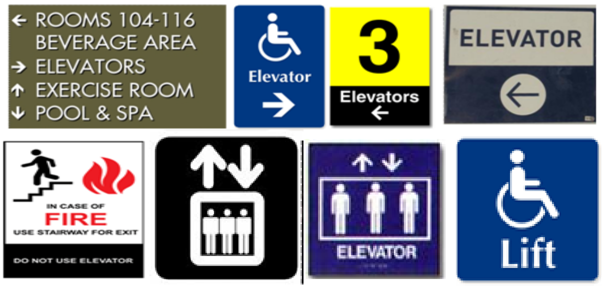


Figure 2. Examples of directional (top) and locational (bottom) signage for elevators.

Figure 2) includes the texts with arrows. The locational signage may include text, a sign, or a combination of text and sign (the second row of Figure 2). For people with normal vision, all of this visual information plays a very important role in finding indoor objects. Other valuable contextual visual information includes buttons and floor numbers etc for finding elevators. The contextual information will guide the model decomposition process for diffident partial-capture situations or diffident object detections. The text on signage is detected and extracted from different video frames. After binarization, the text is recognized by off-the-shelf OCR software.

2. Related works

Supported by psychophysical evidence of the valuable roles that contextual information plays in object detection, how to use context to improve object detection performance is an attractive topic in the computer vision community [8]. Extensive study on context for computer vision has been done by Torralba [10]. However, up to now, there is very little agreement in the literature about the definition of "context." Various visual, no-visual cues are taken as "context" and many methods are reported for combining context into detection.

Zhao and Thorpe [11] proposed a recursive context reasoning (RCR) approach to encode the context information into human object detection. Paletta and Greindl [9] exploited the role of context for system performance in a multi-stage object detection process. They extracted context from simple features to determine regions of interest. Luo *et al.* [7] developed a spatial context-aware object-detection system to improve the accuracy of natural object detection, where the learned spatial context constraints are represented in the form of probability density functions. Very recently, Divvala *et al.* [3] presented an empirical study of the different types of contextual information on a standard, highly regarded test set by incorporating contextual information into a post-process which re-scores detection hypotheses based on their coincidence with the various context cues.

Different from many of the reported context sources, we develop a context-driven decomposition process for our part-based hierarchical compositional model to handle the occlusion problem caused by partial-captured images of the objects. The design of our hierarchical compositional model is inspired by a discriminatively trained and multi-scale deformable part model introduced by Felzenszwalb *et al.* [5]. Unlike the work in [5] using the classic belief propagation and dynamic programming, we adopt the divide and conquer strategy to do inference in a tree-structured graphical model. A compositional hierarchy is defined by breaking down the graphical model into substructures which have their own probability models. The similar idea was reported by Zhu *et al.* [12], however we use a context-driven decomposition process to break down the tree-structured graphical model.

3. Context Information Extraction

3.1. Text Extraction

In order to extract reliable contextual information, we detect the signage and further extract the text on it from different video frames. After binarization, the text is recognized by off-the-shelf optical character recognition (OCR) software.

OCR is considered to be a solved domain for the problem of detecting and reading printed text. However, the success is limited to high quality scanned text images with clean background. Video-based signage detection and text extraction are challenging due to complex color of text and clutter background. Recently, Kasar *et al.* [6] developed a novel technique for binarization of text from digital camera images. It has a good adaptability without the need for manual parameter tuning and can be applied to a broad domain of target document types and environment.

Following the method introduced in [6], edge map E is first obtained by combining the three edge images from three color channels: $E = E_R \vee E_G \vee E_B$. Here, E_R, E_G and E_B are canny edge images from three color channels and $\vee$ denotes the logical OR operation. Then an 8-connected component detection process is employed to assign an edge-box (EB) to each connected region. After filtering out obvious non-text regions by using structure information of text characters, the foreground and background intensities can be estimated by the intensities of edge pixels and their neighbors. Assuming each character has uniform color, we binarize each text character on a clean background by using the estimated foreground intensity as a threshold. Using the binarized text as inputs, a standard OCR software is then able to recognize the text.

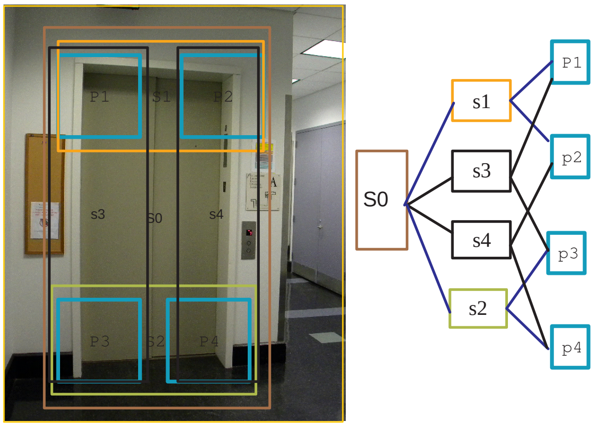
3.2. Distance Estimation

The distance between camera and object plane provides hints for possible partial capture and image quality. From low-cost stereo vision systems, laser sensor based systems, to expensive radar systems, there are different methods which can be used for distance estimation. In this work, we focus on detection of doors which have a standard size range. We estimate the distance between the camera and the object by a simple calibration of the pixels of the object (can be partial of the object) in image with real size of doors.

4. Hierarchical Compositional Object Model

4.1. Model Structure

As shown in Figure 3, an object (e.g. an elevator) is represented by a hierarchical tree-structured graphical model, where the root of the tree S_0 corresponds to the full object which can be decomposed into four parts: Up-part S_1, Bottom-part S_2, Left-part S_3, and Right-part S_4. As shown in Figure 3, parts S_1, S_2 S_3, and S_4 can be further decomposed into child notes: up-left corner p_1, up-right corner p_2, bottom-left corner p_3, and bottom-right corner p_4.



(a) Associated rectangle region of each graphical model.

(b) Tree structure graphical model representation.

Figure 3. Part-based Hierarchical Compositional Model for door detection. Each part is illustrated by a rectangle region.

Similar to a classical tree-structure graphical model [4], each node of this graphical model associates with an appearance model or called filter (see detailed appearance model learning and representation in next subsection.) Each edge between two nodes defines a prior spatial relation which can be assumed to be a Gaussian distribution. Let $l_i^{'} = l_i - l_r$ to be the relative location of a child node s_i to its parent node s_r. Let μ_i and Σ_i be the mean and covariance of the distribution of $l_i^{'}$. These statistical parameters can be obtained by a maximum-likelihood estimator (MLE) from labeled training data. Then, for each child v_i, the conditional distribution of its position is defined below,

$$p(l_i|l_r) = \mathcal{N}(l_i - l_r|\mu_i, \Sigma_i). \qquad (1)$$

An object class is therefore represented by a set of appearance model and geometric distributions. In order to detect objects in different scales, the object detection process is performed by searching over an image pyramid. The features for the part filters $\{f_{p_i}\}, (i = 1, 2, 3, 4)$ are computed at twice the spatial resolution of their next higher-level filters S_1, S_2, S_3, and S_4. The root filter S_0 will operate in the images with the coarsest resolution.

Since there is information redundancy in the model, we employ a context-driven decomposition process to break up the model into more simple subtrees for more effective inference. This is different from a classical tree-structure graphical model [4] which uses the entire graphical model for inference. The detailed context-driven decomposition process is described in the Section 5.

4.2. Appearance Model (filter) Representation

As mentioned before, each graphical model node associates with a part of object within a rectangle image region. For each of these object parts, a linear SVM-based classifier (filter) is trained by using the histogram of oriented gradient (HOG) features.

In training processing, we manually annotated each part with a bounding box. Figure 3(a) shows some samples of image regions within bounding boxes. Following the structure in [2], we define a dense representation of image within each bounding box at a particular resolution. These extracted features from both the positive and negative examples are used to train a linear SVM classifier $f_k(v_{l_i})$, where (k={1,2,...,K}), K is the number of the nodes in the graphical model, v_{l_i} is the HOG feature vector extracted at a slide window position $l_i = (x_i, y_i)$. For a binary classification,

$$f_k(v_{l_i}) = \left\{ \begin{array}{ll} 1, & W_k.v_{l_i} - b_k >= 1; \\ -1, & W_k.v_{l_i} - b_k <= -1, \end{array} \right. \qquad (2)$$

where W_k and b_k are learned SVM parameters for node k. Here we define the value of $f_k(v_{l_i})$ as the distance of a HOG vector v_{l_i} to the SVM hyperplane:

$$W_k.v - b_k = 1$$

$$f_k(v_{l_i}) = W_k.v_{l_i} - b_k - 1, \qquad (3)$$

$f_k(v_{l_i})$ can be simply written as $f_k(l_i)$ in the following formulation. When we conduct this SVM-based classification process in a slide window framework, the function of $f_k(l_i)$ performs as a filter. Given an image I, a map image after filtering is created as: $g_k(I, l_i)$. At each position of l_i in the image I, we have

$$g_k(I, l_i) = f_k(l_i). \qquad (4)$$

5. Context-driven Model Decomposition

As shown in Figure 3(b), the tree-structured graphical model has obvious redundancy information. For example, the Up-part S_1 overlaps with S_3 and S_4; and the Bottom-part S_2 also overlaps with S_3 and S_4. So we never use the full tree-structure graphical model for inference. Instead, we break down this full tree-structured model into one of these star-structured models as shown in the Figure 4 and Figure 5 through a context-driven decomposition process. This decomposition process helps to handle partial captured objects.

5.1. First-layer Decomposition

When the object is far away from the camera, it is less likely to extract text-based information, but more likely to capture images with intact object boundaries. Let d_{is} be the estimated distance between object and camera, when

$$d_{is} > \gamma,$$

where γ is a threshed for d_{is}, no text-based contextual information extraction processing is triggered. At the same time, the full tree-structured graphical model will be decomposed into a subtree as shown in the Figure 4. Object will be represented by a full appearance model S_0 at coarse scale and two part models: Up-part S_1 and Bottom-part S_2 at a finer scale. Considering the information redundancies, part models S_3 and S_4 are pruned.

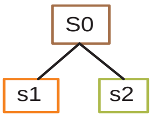


Figure 4. First layer decomposition of the part-based hierarchical compositional model.

5.2. Second-layer Decomposition

When a camera moves closer to the object, it is more difficult to capture images with intact object boundaries. However, higher quality text-based information from signage on/around the object can be extracted. When

$$d_{is} <= \gamma,$$

the process of text-based contextual information extraction is triggered. At the same time, the full tree-structured graphical model will be decomposed into one of the four subtrees as shown in the Figure 5.

In Figure 5, each subtree represents one partial-capture situations, for example, Figure 5(a) represents the Up-part partial-captured case, Figure 5(b) represents the Bottom-part partial-captured case, Figure 5(c) represents the Left-part partial-captured case, and Figure 5(d) represents the Right-part partial-captured case.

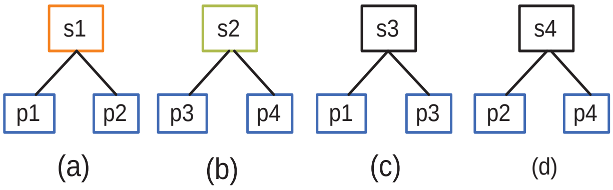


Figure 5. Second layer decomposition of the part-based hierarchical compositional model according to different partial-capture type. (a) Up-part partial-capture. (b) Bottom-part partial-capture. (c) Left-part partial-capture. (d) Right-part partial-capture.

Subtrees in Figure 5(a) and (b) can be used to detect the horizontal object center. Subtrees in (c) and (d) can used to detect the vertical object center. Our investigation shows that, when $d_{is} <= \gamma$, Left-part and Right-part partial captured case seldom happen. Moreover, in order to detect an elevator or a bathroom, the horizontal center of the object provides guidance to navigate the user to move left or right. Therefore, in our algorithm implementation, we omit the case as shown in Figure 5(c) and (d).

In order to understand which subtree and associated filtering processes is used for detection, one approach is to try them all and find the one which has the best fitness to the given image. In our system, given the detected contextual signage, such as floor number or warning signs for elevators, room number or a bathroom sign, let H_{index} denotes the relative position of these contextual signage in an given image,

$$H_{index} = \frac{H - h_y}{H},$$

where H is the height of the given image, h_y is the coordinate of the detected contextual signage in height direction. When a detected contextual signage locates at a high position (with high H_{index} value), it implies that only the Bottom-part of object is captured. Vice versa, a low H_{index} value implies that only the Up-part of object is captured. Let $s_k, k = \{1, 2\}$ be the selected subtree:

$$k = \begin{cases} 1 \text{ (Up-part model)}, & H_{index} < \zeta; \\ 2 \text{ (Bottom-part model)}, & H_{index} >= \zeta, \end{cases} \tag{5}$$

where ζ is a threshed for H_{index} .

6. Combination of Detection and Localization

From Figure 4 and Figure 5 we can see that all these obtained subtrees have the same graphical model structure. Let $G = (V, E)$ be the star-structured graphical model with central node v_r. Given all d node appearance models defined in Equation (3): $A = \{f_k(l_i)|i = 1, ..., n\}$, the intra-class object variation is captured by the spatial relationships between these parts (nodes.) The location of the whole object in an image can be represented by a configuration of its parts $L = (l_1, ..., l_i, ..., l_n)$, where $l_i = (x_i, y_i)$ is the coordinate of ith node in a given image. Following an established line of research [1], the joint prior distribution of one configuration L can be written into the following factorization form,

$$p_A(L) = p_A(l_1, ..., l_n) = p_A(l_r) \prod_{i \neq r} p_A(l_i | l_r). \quad (6)$$

In order to take advantage of the fast inference algorithm presented in [1], we assume that $p_A(L)$ is a Gaussian distribution. Therefore the conditional distribution $p_A(l_i|l_r)$ is still a Gaussian distribution as defined in Equation (4).

Given an image I and learned model parameters A, using Bayes' law, the posterior distribution of an object configuration L can be written as,

$$p_A(L|I) \propto p_A(I|L) p_A(L) \quad (7)$$

Here $p_A(L)$ is the joint prior probability as defined in Function (6); $p_A(I|L)$ is the likelihood of seeing image I given a particular configuration L, which can be calculated based on these filtering results of each part $g_k(I, l_i)$ as defined in Function (3):

$$p_A(I|L) = \prod_{k=1}^{k=n} g_k(I, l_i). \quad (8)$$

Assuming an object is present in an image, its configuration should be the one with maximum posterior probability,

$$L^* = \arg\max_L p_A(L|I) = \arg\max_L p_A(I|L) p_A(L). \quad (9)$$

By manipulating the terms in (6) and (8) we have:

$$p_A(L|I) \propto p_A(l_R) g_R(I, l_R) \prod_{k \neq r} p_A(l_i|l_r) g_i(I, l_i). \quad (10)$$

Then object detection and localization can be jointly obtained by the optimization of Function (10). However, the direct evaluation of it has heavy computational load. We use the efficient inference engine proposed in [1] to obtain L^*. Let write the right side of Function (10) as:

$$p_A(I|L^*) p_A(L^*) = M(L^*). \quad (11)$$

A positive detection happens when

$$M(L^*) > \delta,$$

where δ is a threshold chosen by experimental test.

7. Experiments

In this section we present experimental results for elevator detection. The main purpose of our experiment is to validate the performance improvement of the proposed Part-based HCM approach over a single layer graphical model approach (the case of using the first layer HCM only).

7.1. Experimental Setup

We follow the structure in [2] to define a dense representation of an image within each object bounding box at a particular resolution. For training processing, we manually annotated each object with a bounding box for total 319 objects. For the full object model S_0, 64×128 detection window is represented by 7×15 blocks, giving a total of 3780 features per detection window. For each of parts S_1, S_2, S_3, and S_4, the size of detection window is 128×64. For each part $p_1, ..., p_4$, the size of detection window is 64×64 total of 1764 features per detection window. Also note that the algorithm can handle a limited range of scales. In order to detect text-based contextual information, the algorithm runs at different scale ranges.

7.2. Experimental Results

The effectiveness of the proposed approach is evaluated on a database including 300 test images of elevators which are captured from different buildings with variations of views, scales, and occlusions. An object is deemed successfully detected if the overlap between the detected bounding box and the ground truth bounding box is greater than 50%. The proposed Hierarchical Compositional Model approach can achieve 87.5% recall rate. Without Hierarchical Compositional Model, the single graphical model of the first layer can only get 30.8% recall rate due to occlusion (partial-captured input images.) This recall rate depends on an optimal threshold δ (defined in the Section 6) which is chosen by experimental test.

The first row of Figure 6 shows some examples of the successful detections according to our first layer decomposition models. Since no text-based contextual information is employed at this stage, the algorithm is not able to distinguish an elevator from an office door. From images in the second row of Figure 6, we can see that by using only the first layer decomposition model, it is difficult to handle partial-captured (can be also taken as occluded) objects. The third row of Figure 6 shows the elevator detection results according to our second layer decomposition models. Using the second layer decomposition models, objects can be correctly detected and the horizontal center of objects can be localized. Aided by the extracted locational signage of the elevator and text-based contextual information as shown in the last row of Figure 6, our algorithm is able to detect and distinguish elevators from office doors, bathrooms, and exits.

8. Conclusions and Future Work

In this paper, we have presented a part-based HCM approach for door detection by incorporating context information from signage. We have studied the influence of contextual information for object representation and detection in

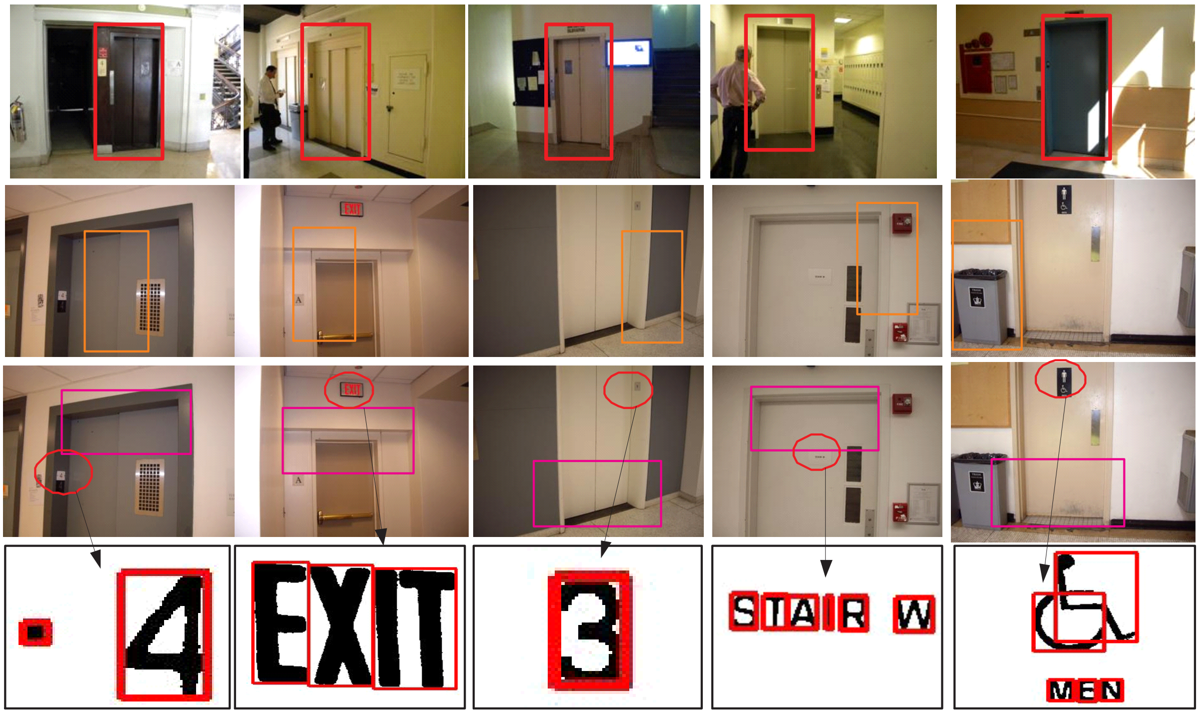


Figure 6. Examples of detection results. The first row shows examples of successful detection results according to the first layer decomposition models. The second row shows failed cases by only using the first layer model but which can be correctly detected by using the second layer models. The fourth row shows the extracted contextual information which help model decomposition

indoor environments. The preliminary results demonstrated that our part-based HCM approach works under variant light conditions and different views, appearance, and scales. The incorporation of contextual information brings significant improvements for partial-captured door detection. We will extend our method to detect more types of indoor objects and improve efficiency of the algorithm in the future. Reliable context information extraction for signage without text information also deserves further investigation.

Acknowledgment

This work was supported by NSF grant IIS-0957016.

References

[1] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. IEEE CVPR*, 2005.

[2] N. Dalal and B.Triggs. Histograms of oriented gradients for human detection. In *Proc. of IEEE CVPR*, 2005.

[3] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proc. of IEEE CVPR*, 2009.

[4] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005.

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. of IEEE CVPR*, 2008.

[6] T. Kasar, J. Kumar, and A. G. Ramakrishnan. Font and background color independent text binarization. In *Second International Workshop on Camera-Based Document Analysis and Recognition*, 2007.

[7] J. Luo, A. Singhal, and W. Zhu. Natural object detection in outdoor scenes based on probabilistic spatial context models. In *Proceedings of International Conference on Multimedia and Expo*, 2003.

[8] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11:520–527, 2007.

[9] L. Paletta and C. Greindl. Context based object detection from video. In *Proceedings of International Conference on Computer Vision Systems*, 2003.

[10] A. Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, 2003.

[11] L. Zhao and C. Thorpe. Recursive context reasoning for human detection and part identification. In *IEEE Workshop on Human Modeling, Analysis, and Synthesis*, June 2000.

[12] L. Zhu and A. Yuille. A hierarchical compositional system for rapid object detection. *In Advances in Neural Information Processing Systems*, 18:1633–1640, 2006.