IEICE TRANS. INF. & SYST., VOL.E92–D, NO.10 OCTOBER 2009

PAPER

Multiple Object Category Detection and Localization Using Generative and Discriminative Models

Dipankar DAS[†a)], **Yoshinori KOBAYASHI**[†], *Nonmembers*, *and* **Yoshinori KUNO**[†], *Member*

SUMMARY This paper proposes an integrated approach to simultaneous detection and localization of multiple object categories using both generative and discriminative models. Our approach consists of first generating a set of hypotheses for each object category using a generative model (pLSA) with a bag of visual words representing each object. Based on the variation of objects within a category, the pLSA model automatically fits to an optimal number of topics. Then, the discriminative part verifies each hypothesis using a multi-class SVM classifier with merging features that combines spatial shape and appearance of an object. In the post-processing stage, environmental context information along with the probabilistic output of the SVM classifier is used to improve the overall performance of the system. Our integrated approach with merging features and context information allows reliable detection and localization of various object categories in the same image. The performance of the proposed framework is evaluated on the various standards (MIT-CSAIL, UIUC, TUD etc.) and the authors' own datasets. In experiments we achieved superior results to some state of the art methods over a number of standard datasets. An extensive experimental evaluation on up to ten diverse object categories over thousands of images demonstrates that our system works for detecting and localizing multiple objects within an image in the presence of cluttered background, substantial occlusion, and significant scale changes.

key words: *object detection and localization, SVM, pLSA, merging feature, context information*

1. Introduction

Multiple object category detection and localization in real, cluttered images is one of the most complex tasks in computer vision. It is critical in many applications such as service robots, image searching, image auto-annotation, and scene understanding. We are currently developing a service robot that can identify an object requested by a user. For this purpose, the robot needs to possess a vision system that can detect and localize various categories of objects in everyday environments. However, this task is still an open problem due to the complexity of objects within an image. Moreover, solving the localization problem requires not only detecting an object, but also determining the precise location of the object within an image. Recent research on object recognition has made great advances with a reasonable recognition rate on many standard datasets. However, most state of the art methods can only solve a binary classification problem [1]–[3]. They are not able to provide information on object location or extent within the image.

Different authors define object localization and detection in different ways. Some techniques define object localization by identifying object parameters [4], [5]. Hierarchical parts-based models giving an estimate of object center as well as its constituent parts have been described in [6], [7]. Some contour segmentation network based approaches have also been described in [8], [9]. However, these seek salient edge groups that are difficult to locate within complex, cluttered backgrounds. Another common approach is to provide a map of the image plane that codes how likely an object is to be presented in a specific pixel [10]. This approach, however, does not explicitly specify the exact location of the object or if there is more than one object present. We have chosen here to localize and detect an object as the placement and evaluation of probable bounding boxes around the object of interest using both generative and discriminative models.

In recent years, the pLSA model with the bag of visual words (BOVW) [2] has been used for categorizing objects because it automatically identifies aspects (topics) from images with semantic meaning. Generative models like pLSA show considerable robustness with respect to partial occlusion, viewpoint, and scale changes [5], [7], [11]. Despite these advantages, the model tends to produce a significant number of false positives. This is particularly true for object classes that share a high visual similarity.

On the other hand, the discriminative method enables the construction of flexible decision boundaries, resulting in classification performance that is often superior to those obtained by purely probabilistic or generative models. Thus, the integration of generative and discriminative models has been used by some authors in order to improve classification performance [16], [17], [27]. However, most state-of-the-art methods have restricted their approach to a single object per image. They are not able to provide information on object locations for multiple objects per image. Moreover, they consider a small number of feature-rich object categories in their experiments. Determining the exact positions of multiple objects within an image is much more difficult than simply saying whether or not there is an object present at all. One has to recognize all objects instead of relying on just the most discriminative ones (as in the case of a single object per image). Also, the image background and other object features in an image are strongly correlated with the presence of certain object features. Perhaps this is the reason why there has been significantly less work done on object detection and localization for multiple objects per image than for a single object per image.

Manuscript received November 25, 2008.
Manuscript revised June 15, 2009.

[†]The authors are with the Department of Information and Computer Sciences, Saitama University, Saitama-shi, 338–8570 Japan.
a) E-mail: dipankar@cv.ics.saitama-u.ac.jp

DOI: 10.1587/transinf.E92.D.2112

In this paper, we improve the integrated approach of the generative and discriminative models to detect and localize multiple objects of various categories by introducing an efficient hypothesis generation method and using an appropriate combination of features. Our approach differs from existing ones both in the generative and discriminative levels. In the generative stage, the pLSA model is fitted to the training data without knowledge of labels of bounding boxes, and topics are assigned based on the image specific topic probability under each category. In our flexible learning strategy, a single object category can be represented with multiple topics and the model can be adapted to diverse object categories with large appearance variations. In the testing stage, an algorithm is proposed and implemented to efficiently generate promising hypotheses for multiple object categories with their positions and scales. For this purpose, our algorithm considers the BOVW extracted from the image and the number of sub-topics generated during the learning stage. Then an initial region of interest (ROI) containing all visual words belonging to a topic is defined based on the maximum topic specific word probability and is searched for a final probable object's locations (promising hypotheses). The initial ROIs reduce the search space and speed up the hypothesis generation stage.

Once hypotheses have been generated, a discriminatively trained SVM classifier verifies these hypotheses using merging features. Since the hypothesis generation stage effectively acts as a pre-filter, the discriminant power is applied only where it is needed. Thus, our system is able to detect and localize multiple objects with a large number of categories. Much of the recent research on object category recognition has developed models focused on single feature− either appearance patches or edge fragments [2], [6], [12]. This is not ideal, as some classes cannot be distinguished by single feature type alone. In this paper we investigate the benefits of merging features over one type of feature alone for the discriminative part of the system. Our merging feature is computed using pyramid HOGs (PHOG) and pre-computed (in the generative stage) visual words based on SIFT descriptors. In order to improve the detection and localization performance we use the category-specific weighted merging features and context information in the post-processing stage.

In this paper we present our object recognition method. Then we demonstrate its performance in experiments using ten diverse object categories of everyday objects that service robots may need to handle.

2. Related Work

The most common approach to generic object detection and localization is a sliding window principle [13]–[15]. The method applies a classifier function subsequently to the sub-images within an image and takes the maximum of the classification score as an indication of the presence of an object in this region. However, it is computationally too expensive to evaluate the quality function exhaustively for all of the image sub-regions. In this paper, the system evaluates the quality function for only probable regions and reduces the computational cost. Our method is inspired by [11]. Instead of creating the doublets vocabulary for segmentation and localization, which require too many doublet probabilities to estimate, all the generated hypotheses are evaluated and verified. Although it was shown that the interest point is able to generate nearly accurate hypotheses about localization of objects in the images for a small number of object categories, there is a proportion of visual synonyms and polysemy for relatively large object categories. For images, the visual synonyms represent several visual words describing the same object or object parts. On the other hand, visual polysemy is a word describing several different objects or object parts. For statistical classification both of the above terms are problematic. This is why only statistical text analysis methods alone are often not powerful enough to deal with the visual words. It has been recently shown that combining the power of generative modeling with a discriminative classifier allows us to obtain good localization and categorization [16], [17]. However, the proposed hybrid approach in [16] was mainly used for scene classification and did not provide any location information of the object. On the other hand, in [17] the same feature is used for both generative and discriminative classifiers and is not sufficient enough to distinguish complex object categories with multiple objects per image. Our approach differs from these in using different features and techniques for both generative and discriminative classifiers.

Some studies [18], [19] for object detection and localization have been conducted to improve both accuracy and speed. Stefan and Manuela proposed and evaluated a method that used PCA-SIFT in combination with a clustered voting scheme with reasonable performance [18]. However they typically restricted their experiments to only two objects. Erik and Jochen demonstrated the feasibility of their approach for relatively large datasets reducing the computational cost [19]. However, the performance of their approach is highly dependent upon the object viewpoints. Moreover, both of the above methods were used to detect and localize specific texture rich objects. Our proposed approach goes beyond these with respect to the following: (i) our discriminative classifier uses both shape and appearance features and can detect and localize object categories with little and no texture (ii) we do experiments over some standard datasets to compare the performance of our method with some state of the art recognition frameworks.

3. Overview of the Proposed Approach

Our proposed approach for multiple object detection and localization is shown in Fig. 1. In the training stage, all the labeled training datasets containing multiple objects per image are presented to the system. In the generative part, the pLSA model [20] is learned for multiple topics using the bag of visual words detected on the uniformly sampled points on the object edges. At the same time the SVM classifier is

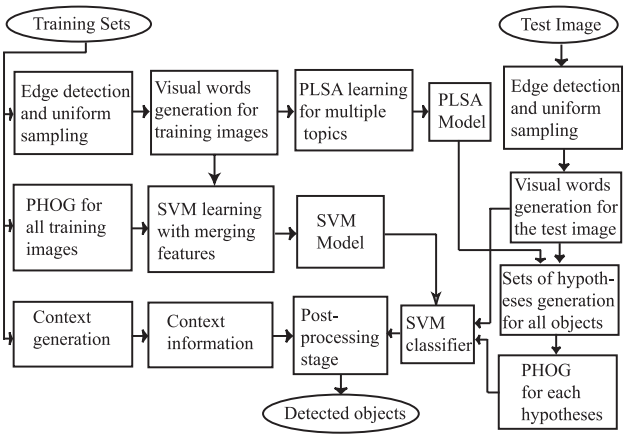


Fig. 1 Fundamental steps of the proposed system.

learned using both a pyramid histogram of oriented gradient (PHOG) and a bag of visual words (BOVW). From the labeled training images we also compute the co-occurrence table to generate the context information. During the testing phase, when a new test image is given, the system generates a set of promising hypotheses with a bag of visual words using the pLSA model. Then we extract the PHOG features from the generated hypotheses and combine them with BOVW features. These merging features and their corresponding locations are verified using the multi-class SVM classifier to detect and localize multiple objects within an image. In the post-processing stage, the generated context information is used with the probabilistic output of the SVM classifier to improve the performance of the system.

4. Hypothesis Generation Using pLSA Model

In order to generate a promising hypotheses for the test image, the pLSA model is first learned using the training datasets.

4.1 pLSA Model

To fit the pLSA model, we first seek vocabulary of visual words for training images that will be insensitive to change in viewpoint, scale, and illumination. These visual words are formed by vector quantizing the SIFT descriptors [21] using the K-means clustering algorithm. The SIFT descriptors are computed on uniformly sampled points in object edges over the circular patch with radius $r = 10$. Uniform sample on the object's edges makes the model shape informative, which is important to get an overall estimate of the object boundary. Therefore, during hypothesis generation, in addition to possible object locations it also gives an estimate of possible object shape. After constructing the visual vocabulary, in the formulation of pLSA for images [11], a co-occurrence table is computed where each image is represented as a collection of visual words. For instance, suppose we have N images containing words from a visual vocabulary of size M. The data is a $M \times N$ co-occurrence table of count $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ stores the number of co-occurrence of word w_i in an image d_j. In addition, there is a latent topic variable $z \in Z = \{z_1, z_2, \ldots, z_K\}$ with each occurrence of a word w_i in an image d_j. The joint probability of $P(w, d, z)$ is defined as $P(w, d, z) = P(w|z)P(z|d)P(d)$. Marginalizing out the latent variable z gives:

$$\begin{aligned} P(w, d) &= \sum_{z \in Z} P(w, d, z) \\ &= P(d) \sum_{z \in Z} P(w|z)P(z|d) \end{aligned} \tag{1}$$

Since $P(w, d) = P(d)P(w|d)$, we obtain $P(w|d)$ as

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{2}$$

Therefore, each image is modeled as a mixture of topics, the histogram for a particular document (image) being composed from a mixture of the histogram corresponding to each topic (object). Here our goal is to determine $P(w|z)$ and $P(z|d)$ by using the maximum likelihood principle with the objective function:

$$L = \log P(D, W) = \sum_{d \in D} \sum_{w \in W} n(w, d) \log P(w, d) \tag{3}$$

The model is fitted for all training images without knowledge of labels of bounding boxes using the Expectation Maximization(EM) algorithm as described in [20] and $P(w, d)$ is given by Eq. (1). Then topics are assigned based on the image specific topic probability under each category.

4.2 The Promising Hypotheses Generation

The pLSA model determines the mixture coefficients $P(z_k|d_j)$ for each object d_j. An object d_j is then classified as to maximum $P(z_k|d_j)$ over the number of topics, k. An object category may belong to multiple sub-topics. When a new test image is given, all visual words are extracted from objects and background in the image and each visual word is classified under the topic with the highest topic specific probability $P(w_i|z_k)$. Then it is used to detect the region of interest (ROI) for each object category in the image. The ROI is the smallest rectangular region within the image that contains all possible visual words for a particular object category. As an example, Fig. 2 (a) shows the original image with four target object categories, namely coffee jar, coffee mug, spoon, and cup noodle. For simplicity, among four detected ROIs Fig. 2 (b) shows one of the ROIs and its corresponding possible visual words. Visual words are drawn in small circles on the image. As shown in this figure, ROIs are generally large because of existence of visual words derived from other objects and background than target objects due to visual polysemy. The following algorithm can efficiently generates promising hypotheses within those ROIs.

1. For all object categories repeat the following steps with their corresponding rectangular ROI.

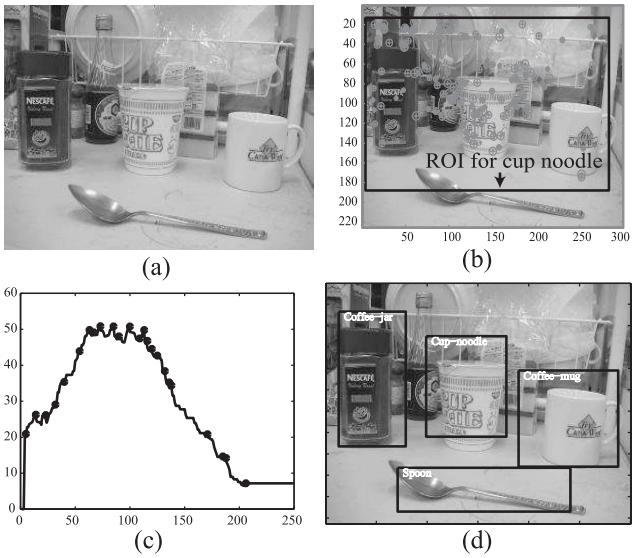


Fig. 2 Hypotheses generation and SVM verification results: (a) original image with four target objects (b) detected ROI for the object *cup noodle* (c) local maxima for *cup noodle* and (d) detected target objects with their extents.

2. Compute the average aspect ratio, M_{a_i} of the window for each object category i as $M_{a_i} = M_{w_i}/M_{h_i}$, where M_{w_i} and M_{h_i} are mean width and height of the object i computed during the training stage using ground truth bounding boxes.
3. For each object category, slide the window with the average aspect ratio, M_{a_i} and count the number of visual words, $N_{vw} = \sum_{z \in t_s} n_{vw_{iz}}$, where $n_{vw_{iz}}$ is the number of visual words for object category i and sub-topics t_s. A category may belong to one or more sub-topics.
4. Determine the local maxima (Fig. 2 (c)) based on the average number of visual words at each column position calculated as: $N_{avg} = \frac{1}{R}\sum_{r=1}^{R} N_{vw}$ where R is the number of rows for which sliding window repeats within ROI.
5. For all local maxima regions within an image find and suppress the windows, if any, which overlap by 75% or more with the window that contains the maximum number of visual words for each local region. This step is almost similar to the non-maximum suppression technique.
6. After suppressing the non-maximum windows in each neighborhood the remaining windows are selected as the promising hypotheses.

5. SVM Verification with Merging Features

It has been shown in [16] that pLSA provides a better intermediate representation of images using a bag of visual words. On the other hand, object detection and localization algorithms that use discriminative methods combined with global and/or local representation have been shown to perform well in the presence of clutter, viewpoint changes, partial occlusion, and scale variations [22]. In our approach, along with pLSA, a multi-class support vector machine (SVM) classifier is also learned in parallel using shape and appearance features. To represent the shape of an object, spatial shape descriptors are extracted from the object of interest. In order to describe the spatial shape of an object we follow the scheme proposed by Anna Bosch *et al.* [14]. Here the object is represented by its local shape and spatial layout. The local shape is represented by orientations of an edge histogram within an object's subregion quantized into K-bin and each edge's contribution is weighted by its magnitude. Therefore, each bin in the histogram represents the number of edges that have orientations within a given angular range. The spatial layout is given by tiling the object into regions at multiple resolutions. As a result, the final shape descriptors consist of a histogram of orientation gradients over each object sub-region and at each resolution level− a Pyramid Histogram of Orientation Gradient(PHOG). The final shape descriptor of the entire object is a vector with dimensionality $K\sum_{l \in L} 4^l$ and is normalized to sum to unity so that some objects (edge rich) are not weighted more strongly than others. Although shape representation is a good measure of object similarity for some objects (e.g. coffee mug, CD), shape features are not sufficient enough to distinguish among all types of objects (e.g. keyboard, book). In this case, object appearance represented by the bag of visual words is a better feature to find the similarity between them. The appearance patches and descriptors are computed in a similar manner as described in Sect. 4. Then the normalized histogram of visual words for each object is calculated. Finally, the combination of both shape and appearance features for an object O, are merged as:

$$H(O) = \alpha H_S(O) + \beta H_A(O) \tag{4}$$

where both α and β are weights for the shape histogram, $H_S(O)$ and appearance histogram, $H_A(O)$, respectively. The multi-class SVM classifier is learned using the above merged feature giving the higher weight to the more discriminative feature. The values of α and β in Eq. (4) are determined for each object category separately on the cross-validation dataset. We use the LIBSVM [23] package for our experiments in a multi-class mode with the *rbf* exponential kernel.

In the verification step, merging features are extracted from regions of the image bounded by windows of the promising hypotheses. For all windows in the test image, shape and appearance descriptors are combined according to the Eq. (4) and fed into the multi-class SVM classifier in recognition mode. Only the hypotheses for which a positive confidence measurement is returned are kept for each object. Objects with the highest confidence level are detected as the correct objects, (Fig. 2 (d)). The confidence level is measured using the probabilistic output of the SVM classifier. Our SVM verification stage is very fast because only a few locations per image need to be verified. In the testing phase, visual words generation takes a longer time due to calculation of SIFT descriptors. However, it only needs to be done once, so the cost is amortized when searching for

windows of all object categories.

6. Post-Processing Using Context Information

In the task of object category recognition, environmental context information can play an important role of reducing ambiguity in an object's visual appearance. Rabinovich et al. [24] used a semantic context of inter-class dependencies to improve object detection performance. Their method segments the image and classifies all segments jointly based on a conditional random field. However, this requires the object dependencies to be specified a priori, whereas our method learns environment-related context information during the training process and uses this information along with probabilistic output of the SVM classifier to improve object detection and localization performance. To incorporate context information in our system, we first construct context matrices. These are symmetric, non-negative matrices that contain co-occurrence frequency among object labels in the training sets of the database. Then fully connected context graphs are constructed from these co-occurrence matrices. Thus, a separate context graph is built for every environmental dataset in our experiment. Edges and vertices of the graph are represented by the co-occurrence frequencies and object categories, respectively. During post-processing stage, first the base context is determined by using the output of the SVM classifier. For this purpose, both number of detected objects and their probabilities are used. A context graph that belongs to the maximum number of detected objects is selected as a base context. If the number of detected objects are equal for multiple context graphs, then the context graph that belongs to the maximum total probability of the detected objects is used to select it. The base context information is then used to give the flexible margin for the context-related (the relation is determined using the context graph) objects and hard margin for non-contextual objects. In some cases, the context information incrementally increases the false positive rate for intra-contextual objects. However, it decreases the overall false positive rate and increases the overall detection rate. In this research, it is mainly used to improve the detected performance of the SVM classifier. For example, suppose that in an office environment, an image consists of five target objects (book, CD, computer monitor, computer keyboard, and computer mouse) as shown in Fig. 3 (a). In order to verify the presence of an object and its extent, a threshold margin is set for the probabilistic output of the SVM classifier. Without context information this threshold margin is set to 0.4 for the ten object categories in the experimental evaluation on our dataset. In this case, the system detects four target objects (book, computer monitor, computer keyboard, and computer mouse) and two false objects (coffee mug and cup noodle). The other object CD is detected with a low confidence level (0.37) and is not included in our intermediate result (Fig. 3 (b)). However, using the context information the threshold margin is set to 0.75 for non-contextual objects (coffee mug and cup noodle). On the other hand, if an object is detected correctly then no restriction is set for context-related object (CD). The relationship of objects with the base context is determined by using the context graph. Figure 3 (c) shows the final result of using context in the pos t-process ing stage.

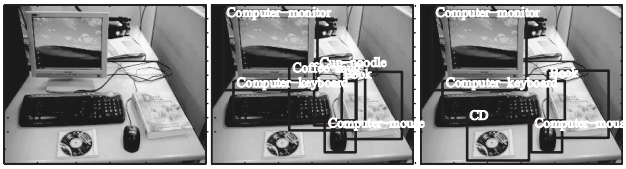


Fig. 3 Performance improvements using context: (a) original image with five target objects (b) detected objects without context and (d) detected objects with context.

7. Datasets

We evaluate our detection and localization algorithm on different datasets. In order to compare our approach with [5], [17], [25] the same four categories of objects, namely cars from UIUC and TUD car datasets, cows from TUD cow dataset, horses from Weizmann horse dataset and motorbikes from CalTech motorbike dataset, are used. Most of these datasets are taken from the PASCAL database collection [26] and contain a single object per image. For multiple objects per image the performance of the system is evaluated on two different datasets: the MIT-CSAIL static office dataset and our own dataset. The MIT-CSAIL office datasets are collected from PASCAL VOC 2007 database collection for three categories of objects: computer monitors, computer keyboards, and bookshelves. The datasets contain multiple objects per image and these categories were selected because they occurred more frequently in images. From different static office datasets a total of 125 images containing 252 objects are used for our experiment. Among them the training dataset contains 65 images of 133 annotated objects. In the remaining 60 images, a total of 119 objects are used for testing purposes. Finally, we evaluate the performance of the system on our own datasets. It consists of 10 categories of everyday objects related to our application (service robot) in different environments against cluttered, real-world background with occlusion, scale, and minor viewpoint changes. Our datasets are created with ground truth bounding boxes that contain multiple objects per image. There are a total of 774 images containing 2002 objects. Among them 293 images (with 582 objects) are used for training purposes and the rest of the 481 images (with 1420 objects) are used for testing. Since objects were presented randomly within an image, there exist differences in depth, position, rotation, and lighting. The depth changes caused a significant amount of scale variation among objects. Ten categories of objects were grouped into two datasets. Dataset-1 consists of five categories of kitchen environment objects, namely coffee jar, coffee mug, spoon, hand soap, and cup noodle. Dataset-2 consists of

another five categories of office environment objects: computer monitor, computer keyboard, computer mouse, CD, and book.

8. Experimental Results

In this section we carry out a set of experiments to investigate the benefits of our integrated approach with merging features and context information. Given a completely unlabeled image of multiple object categories, our goal is to automatically detect and localize objects in the image. In our approach, object presence detection means determining if one or more categories of objects are present in an image and localization means finding the location of an object in that image. Based on the object presence detection and localization, an object is counted as a true positive object if the detected object boundary overlaps by 50% or more with the ground truth-bounding box for that object. Otherwise, the detected object is counted as false positive. In the experimental evaluation, the detection and localization rate (DLR) is defined as:

$$\text{DLR} = \frac{\text{\# of true positive objects}}{\text{\# of annotated objects}} \tag{5}$$

The false positive rate (FPR) is defined as:

$$\text{FPR} = \frac{\text{\# of false positive objects}}{\text{\# of annotated objects}} \tag{6}$$

To understand how the proposed method performs, in the following subsections we investigate four areas: parameter optimization, comparisons with other methods, benefits of the integrated method, and performance on MIT-CSAIL and the authors' own datasets. For simplicity, we start by finding the optimal parameter values on our own dataset. For the rest of the experiments we apply similar techniques to find the optimal parameter values.

8.1 Parameter Optimization

In this section, we investigate how detection and localization performance is affected by the various parameters: number of visual words (W), number of topics (Z), values of α and β, cost parameter (C) and kernel parameter (γ) of SVM, and threshold margins of post-processing stage.

The number of visual words (W) and the number of topics (Z) directly affect the performance of the accurately generated hypotheses. Thus, we optimize the parameters W and Z on a validation dataset based on the performance of the hypotheses generation accuracy. Our validation dataset consists of 104 images of 307 objects. Figure 4 shows the performance variation for two parameters W and Z. The best performance is obtained with $W = 600$ and $Z = 20$ as indicated by the circular markers on the graph. With $Z = 20$, the pLSA model sub-divides the object and background categories as shown in the Table 1.

Since the values of α, β, C, and γ are directly related to the recognition results of the SVM classifier, we determine these values during a training period using five-fold cross validation ($v = 5$). We first determine the values of C and γ using the grid search technique of the SVM classifier. The best values of C and γ for our experiments are 66 and 0.5, respectively. Then the pairs (α, β) are tried for each category separately and the one with the best cross-validation accuracy is picked. In this case, the values α and β are varied within the range 0 to 1 by step 0.5. The example values of (α, β) for different categories of objects with final cross validation accuracy of 98.97% are shown in Table 2.

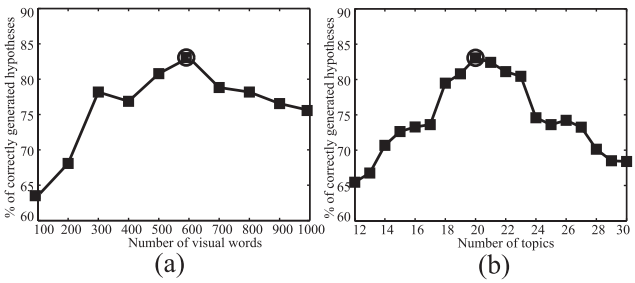


Fig. 4 Validation set performance under variation in parameters for the ten category authors' dataset: (a) performance vs W ($Z = 20$), (b) performance vs Z ($W = 600$).

The threshold margins of the post-processing stage are determined using the validation dataset. Without context, the threshold margin of 0.4 is selected as the lowest probability among correctly detected objects on the validation dataset. With context, the threshold margin of 0.75 is selected as the average probability of all correctly detected objects on the same dataset.

8.2 Comparison with Other Methods

The performance of our system is compared to the integrated representative and discriminative (IRD) representation of Fritz et al. [17], the implicit shape model (ISM) of Leibe et al. [5] and local kernels (LK) representation of Wallraven et al. [25], using the same datasets that are tested in [17]. For each dataset we use the SVM classifier with PHOG feature to verify the hypotheses generated by our algorithm as discussed in Sect. 4. Table 3 summarizes the performance of our experiment with other methods. The test is performed on images of each category versus 200 Caltech-101 and Caltech-256 background images. Each image in these datasets contains only one target object as shown in Fig. 5. Although the recognition task is different from our multiple object detection and localization, we performed this experiments to compare basic performance of our method with others.

Note that in the majority of cases better results are obtained with three categories of objects. The classification result of our approach on UIUC car dataset is slightly less than [17]. However, the result is better than the other two methods. We also evaluated the performance of our method on TUD car datasets and obtained recognition accuracy of 98.3%. The superior performance compared to [17] could

Table 1 Example sub-categories with $Z = 20$ for the authors' dataset.

Category	Coffee jar	Coffee mug	Spoon	Hand soap	Cup noodle	Monitor	Keyboard	Mouse	CD	Book	Back-ground
#of topics	2	2	1	2	1	3	2	1	2	2	2

Table 2 The values of α and β with five-fold cross validation ($v = 5$) accuracy 98.97%.

Category	Coffee jar	Coffee mug	Spoon	Hand soap	Cup noodle	Monitor	Keyboard	Mouse	CD	Book
α	1	1	1	1	0.5	1	0.5	1	1	0.5
β	1	0.5	1	1	1	1	1	0.5	0.5	1

Table 3 Performance comparison of our algorithm with other methods.

Category and Dataset	LK [25]	ISM [5]	IRD [17]	Authors
Horse (Weizmann)	77.8%	88.5%	88.5%	**97.0**%
Cow (TUD)	95.3%	96.1%	97.1%	**98.6**%
Motorbike (CalTech)	87.6%	93.8%	96.5%	**98.3**%
Car (UIUC)	61.0%	94.7%	99.4%	97.1%
Car (TUD)	–	–	–	98.3%

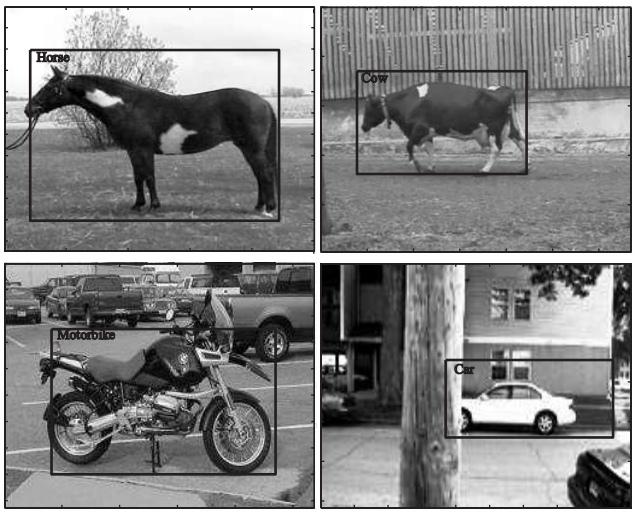


Fig. 5 Detection results on the horse, cow, motorbike and car datasets.

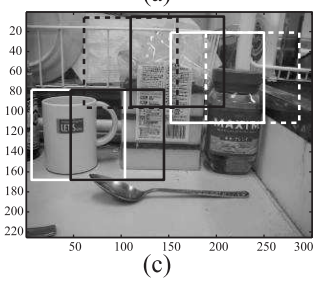
be due to the use of better features and how they are used in our approach. In [17], they used the same feature for both generative and discriminative classifiers. In our approach different features are used for both generative and discriminative parts. Example detection and localization results on different datasets are shown in Fig. 5.

8.3 Benefits of the Integrated Method

In this section we will investigate the benefits of our SVM verification stage instead of using only pLSA for detection and localization purpose. As we previously mentioned, the generative model alone is not sufficient enough to detect multiple objects in an image. This is due to visual polysemy. The problem becomes apparent when we consider how an image is represented in the bag of visual words documents model. All visual words in an object are represented by a single histogram, and lose all spatial and neighborhood relationships. In the following experiments, we use our own

Number of visual words	Window size Left top		Bottom right	
21	6	78	104	168
19	189	21	287	111
15	152	21	250	111
9	47	78	145	168
7	110	5	208	95
5	61	6	159	96

(b)



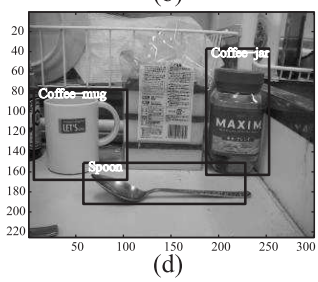


Fig. 6 Results of the integrated method: (a) the original test image, (b) number of visual words and window size for the object, *coffee mug*, (c) windows for the object, *coffee mug* on the image and, (d) verified *coffee mug* object along with other two objects by SVM classifier.

dataset containing four object categories: coffee jar, coffee mug, spoon, and hand soap. The training and testing datasets consist of 111 images of 160 objects and 130 images of 420 objects, respectively. In our experimental result, let us consider the original image of Fig. 6 (a). In this case, the number of visual words generated for *coffee mug* object in different probable windows is given in Fig. 6 (b) and their corresponding regions of windows are shown in Fig. 6 (c). From the illustration it is clear that a significant amount of visual words are generated from the other areas than the *coffee mug* object due to the visual polysemy nature of objects and/or object parts and complex backgrounds. However, there is strong evidence among the generated hypotheses for the *coffee mug* object in the image and is verified by the SVM classifier. Figure 6 (d) shows the final detected results of our integrated method for the *coffee mug* object along with two other objects(*coffee jar and spoon*).

In this section, we also investigate how our pLSA method performs for generating the promising hypotheses. Table 4 shows the detected objects by our hypotheses generation method, where $w_i, i = 1 \ldots 7$, indicates the correctly detected hypothesis window. Using only pLSA, if we take the maximum number of visual words that belong to w_1 win-

Table 4 Hypotheses generation and SVM verification results.

Category	Detected objects							Undetected object	SVM verification results
	w_1	w_2	w_3	w_4	w_5	w_6	w_7		
Coffee jar	65	28	8	3	–	–	–	2	90
Coffee mug	9	11	18	11	16	11	9	26	76
Spoon	62	23	3	–	–	–	–	2	90
Hand-soap	20	23	19	14	12	12	9	2	89
Avg. (%)	**37**	**20**	**11**	**7**	**7**	**6**	**4**	**8**	**80**

dow for classification purposes then only 37% objects are detected. Similarly, the window containing a second maximum number of visual words (w_2) detects only 20% of the total numbers of objects, and so on. However, from Table 4 it is clear that all of the generated hypotheses are able to detect 92% objects. Using the SVM verification stage on the generated hypotheses our system detects 80% of total objects as shown in the last column of Table 4.

8.4 Performance on MIT-CSAIL and Authors' Datasets

In this section we measure the detection and localization performance of our approach on MIT-CSAIL and our own datasets. In the training period, both the pLSA and SVM models are fitted for all training object categories. The values of different parameters are obtained over validation and cross-validation datasets using the technique as described in Sect. 8.1.

8.4.1 Results—— MIT-CSAIL Dataset ——

Table 5 summarizes the performance achieved on the MIT-CSAIL static office datasets for three categories of objects. Experiments were done using the shape feature, appearance feature, and finally a combination of both of them. As can be seen in this table, poor performance is obtained for the dataset when we use only the shape or only the appearance feature. However, there is a significant improvement in the results when merging features and object specific weighted merging features are used. The average detection and localization result for the weighted merging feature is 82% with a false positive rate of 0.29. Weighted merging features increase the average detection and localization rate by 16% compared to appearance features alone (66%). Figure 7 shows two examples of detection and localization results on this dataset. Our final result is comparable with Sivic et al. [11] for some categories of objects. In their approach, 15 out of 20 computer screen (75%) and 17 out of 20 bookshelves (85%) are are correctly detected. On the other hand, in our approach the detection and localization accuracy for computer screen and bookshelf are 84% and 93%, respectively. Our better performance compared to [11] could be due to the integration of both generative and discriminative classifiers instead of using only generative model.

Table 5 Experimental results on MIT-CSAIL dataset.

Category	Appearance feature		Shape feature		Merging feature		Weighted merging feature	
	DLR	FPR	DLR	FPR	DLR	FPR	DLR	FPR
Monitor	0.67	0.12	0.71	0.06	0.78	0.08	0.84	0.14
Keyboard	0.60	0.08	0.68	0.04	0.70	0.09	0.77	0.11
Bookshelf	0.87	1.00	0.87	2.20	0.93	1.80	0.93	1.47
Avg. Rate	**0.66**	**0.21**	**0.71**	**0.32**	**0.76**	**0.30**	**0.82**	**0.29**

Fig. 7 Detection and localization results on MIT-CSAIL.

8.4.2 Results—— Authors' Dataset ——

In a series of experimental evaluations, we finally evaluate the performance of the system in our own dataset. In most of the studies [11], [17], [18], [22], a small number of categories (two to five) were used for categorization purposes. Thus, we collected the dataset consisting of ten categories of objects in different environments and backgrounds. A detailed description of the dataset is given in Sect. 7. Table 6 shows the detection and localization rate at the false positive rate indicated in their adjacent column. In this experiments, during the training period both the pLSA and SVM model are fitted for all ten categories of objects. Based on the variation of objects within a category our pLSA model automatically fits for multiple topics. For the experiments on our dataset the pLSA model is fitted for 20 topics (two topics for background and the rest 18 topics for 10 object categories). Images with multiple objects along with their ground truth bounding boxes are used to determine the context information (a matrix of label co-occurrence count). Since the merging feature outperforms the single feature, we have used the merging feature with weight and context information for this experiment. As shown in Table 6, the merging feature without any weight and context information produces an average DLR 66%. On the other hand, using the same feature with context information as a post-processing stage, the system incrementally increases the average DLR to 68% with a reduction of the false positive rate from 37% to 24%. Since some objects are best described by their shape feature (e.g. coffee mug, CD) and others by their appearance (e.g. computer keyboard, book), the weighted merging feature gives us the best performance (77%) for all ten object categories. Although the context information incrementally increases the detection and localization performance, it significantly decreases the false positive rate. Some detection

Table 6 Experimental results on authors' datasets.

Category	Merging feature		Merging feature with context		Weighted merging feature		Weighted merging feature with context	
	DLR	FPR	DLR	FPR	DLR	FPR	DLR	FPR
Coffee jar	0.81	0.12	0.81	0.09	0.80	0.22	0.81	0.13
Coffee mug	0.30	0.09	0.31	0.05	0.58	0.74	0.73	0.40
Spoon	0.76	0.08	0.80	0.19	0.75	0.11	0.78	0.08
Hand soap	0.55	0.12	0.55	0.11	0.70	0.56	0.74	0.28
Cup noodle	0.81	0.75	0.83	0.48	0.79	1.11	0.84	0.54
Monitor	0.80	0.08	0.81	0.05	0.86	0.25	0.88	0.03
Keyboard	0.90	0.11	0.90	0.07	0.97	0.21	0.97	0.05
Mouse	0.60	0.91	0.60	0.95	0.60	1.44	0.63	0.67
CD	0.46	0.83	0.47	0.26	0.58	1.31	0.58	0.50
Book	0.63	1.80	0.68	0.91	0.64	1.91	0.68	0.76
Avg. Rate	**0.66**	**0.37**	**0.68**	**0.24**	**0.73**	**0.68**	**0.77**	**0.30**

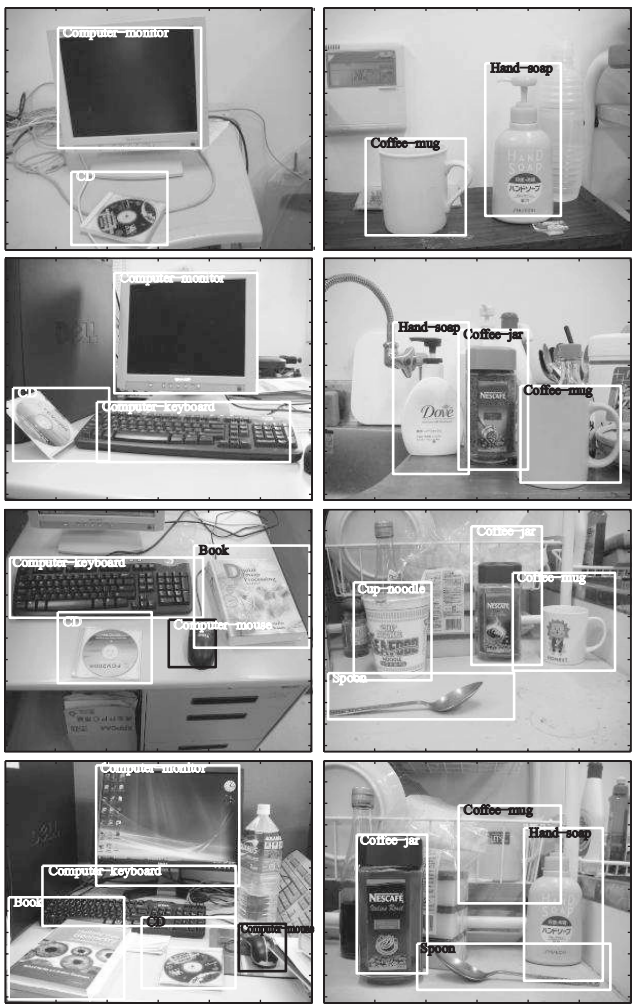


Fig. 8 Example detection and localization results on authors' datasets.

and localization results on our own datasets are shown in Fig. 8. Based on object appearance and viewpoint changes our system sub-categorizes an object category into one, two or three sub-topics. Thus the system is able to detect objects with minor viewpoint changes as shown in Fig. 7 and Fig. 9.

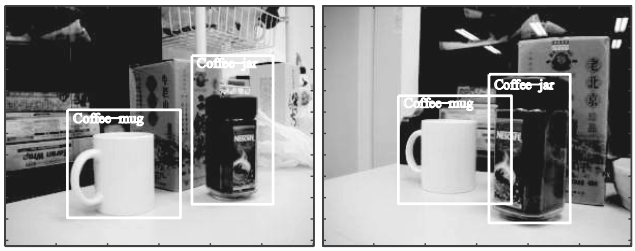


Fig. 9 Detected image region with viewpoint change.

9. Conclusion

In this paper, we have proposed an approach of integrating both the generative and discriminative classifiers for multiple object category detection and localization. Our system has shown the ability to accurately detect and localize many objects even in the presence of cluttered background, substantial occlusion, and significant scale changes. Our experimental results demonstrate that the hypotheses generation algorithm is able to generate nearly accurate hypotheses for all object categories. The SVM verification stage, on the other hand, uses the merging features and category specific weighted merging features to enrich the performance of the system. Finally, the environmental context information in the post-processing stage compensates for ambiguity in an object's visual appearance.

One of the fundamental problems in 3D object recognition how to deal with object appearance changes depending on the viewpoint. The current system can handle small viewpoint changes. And, in theory, if the training data consist of images with large viewpoint changes, then the generative model automatically subcategorize objects in a given category into multiple topics and generate nearly accurate hypotheses. However, to do this, we need to show many images from various viewpoints. And the number of subcategories may increase greatly. We are now working on how to deal with this problem. We will explore the possibility of detecting objects with large viewpoint changes by automatically sub-categorizing the object categories into the appropriate number and using the variable size object windows based on detected visual words. We also plan to use the environmental context information in more meaningful ways to detect and localize missing objects within an image depending on the base context environment.

Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research (KAKENHI 19300055).

References

[1] A. Mansur and Y. Kuno, "Specific and class object recognition for service robots through autonomous and interactive methods," IEICE

Trans. Inf. & Syst., vol.E91-D, no.6, pp.1793–1803, June 2008.
[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Proc. European Conf. on Computer Vision (ECCV'04), Workshop on Statistical Learning in Computer Vision, Prague, 2004.
[3] A. Diplaros, T. Gevers, and I. Patras, "Combining color and shape information for illumination-viewpoint invariant object recognition," IEEE Trans. Image Process., vol.15, no.1, pp.1–11, 2006.
[4] X. Stella, G. Ralph, and S. Jianbo, "Concurrent object recognition and segmentation by graph partioning," Proc. Neural Information Processing Systems (NIPS), pp.1383–1390, Vancouver, Canada, 2002.
[5] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," Proc. ECCV'04, Workshop on Statistical Learning in Computer Vision, pp.17–32, Prague, 2004.
[6] B. Guillaume and T. Bill, "Hierarchical part-based visual object categorization," Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR(1)), pp.710–715, San Diego, CA, USA, 2005.
[7] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of model for visual recognition," Int. J. Computer Vision (IJCV), vol.71, no.3, pp.273–303, 2007.
[8] V. Ferrari, T. Tinne, and V.G. Luc, "Object detection by contour segmentation networks," Proc. ECCV(3), pp.14–28, Graz, Austria, 2006.
[9] D. Jacobs, "Robust and efficient detection of salient convex groups," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol.18, no.1, pp.23–37, 1996.
[10] M. Marcin and S. Cordelia, "Spatial weigthing for bag-of-features," Proc. CVPR(2), pp.2118–2125, New York, NY, 2006.
[11] S. Josef, C. Bryan, Russell, A. Alexei, A. Zisserman, and T. William, "Discovering objects and their location in images," Proc. IEEE Int. Conf. on Computer Vision (ICCV'05), pp.370–377, Beijing, China, 2005.
[12] A. Opelt, A. Pinz, and A. Zisserman, "A boundary-fragment-model for object detection," Proc. ECCV (2), pp.575–588, 2006.
[13] O. Chum and A. Zisserman, "An exemplar model for learning object classes," Proc. CVPR, IEEE Computer Society, 2007.
[14] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with spatial pyramid kernel," Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR), pp.401–408, Amsterdam, The Netherlands, 2007.
[15] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Group of adjacent contour segment for object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.1, pp.30–51, 2008.
[16] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.4, pp.712–727, 2008.
[17] M. Fritz, B. Leibe, B. Caputo, and B. Schiele, "Integrating representative and discriminative models for object category detection," Proc. ICCV'05, Beijing, China, pp.1363–1370, 2005.
[18] Z. Stefan and M. Manuela, "Detection and localization of multiple objects," Proc. Humanoids, Genoa, Italy, 2006.
[19] M.C. Erik and T. Jochen, "Shared features for scalable appearance-based object recognition," Proc. IEEE Workshop on Application of Computer Vision(WACV), pp.16–21, Breckenridge, Colorado, 2005.
[20] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Mach. Learn., vol.42, no.1/2, pp.177–196, 2001.
[21] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis. (IJCV), vol.60, no.2, pp.91–110, 2004.
[22] K.P. Murphy, A.B. Torralba, D. Eaton, and W.T. Freeman, "Object detection and localization using local and global features," Toward Category-Level Object Recognition, pp.382–400, Springer, 2006.
[23] C.C. Chang and C.J. Lin, "Libsvm: A library for support vector machines." http://www.csie.ntu.edu.tw/cjlin/libsvm/, 2008.
[24] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," Proc. ICCV'07, Rio de Janeiro, Brazil, 2007.
[25] C. Wallraven, B. Caputo, and A.B.A. Graf, "Recognition with local features: The kernel recipe," Proc. ICCV, pp.257–264, IEEE Computer Society, 2003.
[26] "The PASCAL Visual Object Classes." http://www.pascal-network.org/challenges/VOC
[27] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," Proc. NIPS, pp.487–493, Denver, Colorado, USA, 1998.

Dipankar Das received his B.Sc. and M.Sc. degrees in Computer Science and Technology from the University of Rajshahi, Rajshahi, Bangladesh in 1996 and 1997, respectively. He is an assistant professor of the Department of Information and Communication Engineering of the same university. Currently, he is also a PhD student in the Graduate School of Science and Engineering, Saitama University. His research interests include Object Recognition, Human Computer Interaction and Speech Processing.

Yoshinori Kobayashi completed the M.E. degree from the Department of Information Management Science at Graduate School of Information Systems, University of Electro-Communi-cations in 2000, and joined the Design Systems Engineering Center of Mitsubishi Electric Corporation. He was in the doctoral program in 2004-2007 in Information and Communication Engineering at the Graduate School of Information Science and Technology, The University of Tokyo, and then he joined the Department of Information and Computer Sciences, Saitama University, as an assistant professor. He is interested in computer vision for human sensing and its application for human computer interaction.

Yoshinori Kuno received B.S., M.S. and PhD degrees in 1977, 1979 and 1982, respectively, all in Electrical and Electronics Engineering from The University of Tokyo. In 1982, he joined Toshiba Corporation. From 1987 to 1988, he was a Visiting Scientist at Carnegie Mellon University. In 1993, he moved to Osaka University as an associate professor in the Department of Computer Controlled Mechanical Systems. Since 2000, he has been a professor in the Department of Information and Computer Sciences, Saitama University.