An Integrated Method for Multiple Object Detection and Localization

Dipankar Das, Al Mansur, Yoshinori Kobayashi, and Yoshinori Kuno

Graduate School of Science and Engineering, Saitama University,
255 Shimo-Okubo, Sakura-ku, Saitama-shi, Saitama 338-8570, Japan
{dipankar,mansur,yosinori,kuno}@cv.ics.saitama-u.ac.jp

Abstract. The objective of this paper is to use computer vision to detect and localize multiple object within an image in the presence of a cluttered background, substantial occlusion and significant scale changes. Our approach consists of first generating a set of hypotheses for each object using a generative model (pLSA) with a bag of visual words representing each image. Then, the discriminative part verifies each hypothesis using a multi-class SVM classifier with merging features that combines both spatial shape and color appearance of an object. In the post-processing stage, environmental context information is used to improve the performance of the system. A combination of features and context information are used to investigate the performance on our local database. The best performance is obtained using object-specific weighted merging features and the context information. Our approach overcomes the limitations of some state of the art methods.

1 Introduction

Object detection and localization is an important yet challenging task in computer vision, especially in the presence of pose changes, occlusion and background clutter. 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 objects in ordinary environments. However, it 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. Recently object recognition from images has made a lot of advances with a reasonable recognition rate on many datasets. However, most state of the art methods can only solve a binary classification problem[1,2,3]. They are not able to provide information on object locations or extent within the image.

Locating the exact position of an object within an image is much more difficult than simply saying whether there is an object present or not. Different authors define object localization and detection in different ways. Some techniques define object localization by identifying objects parameters[4,5]. Hierarchical parts based models giving an estimate of object center as well as its

G. Bebis et al. (Eds.): ISVC 2008, Part II, LNCS 5359, pp. 133–144, 2008.

constituent parts have been described in[6,7]. Some contour segmentation network based approaches are described in[8,9]. However, they seek salient edge groups that are very difficult to locate within complex, cluttered background. 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]. The approach does not explicitly specify the exact object location 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 most probable bounding boxes around the object of interest using both generative and discriminative models.

Sliding window bounding boxes have been used extensively in the field of object localization[11,12]. The sliding window principle applies a classifier function subsequently to the subimages within an image and takes the maximum of the classification score as indication for the presence of an object in this region. It is computationally too expensive to evaluate the quality function exhaustively for all of the image subregions. In this paper, we propose an integrated method to perform an object detection and localization in a way that alleviates these drawbacks. The method is based on finding one or more possible object locations within an image using a generative model and then evaluating these locations using a discriminative classifier. We have two goals: the first is to find possible object locations and scales (hypotheses generation). The hypotheses are generated by counting the number of individual visual words within a window that belong to a particular object and then finding the most probable windows for that object. The second goal is to evaluate the hypotheses by a classifier that uses both appearance and shape features of the object. This significantly reduces the computational time because it requires evaluating only few locations per image.

Our method is inspired by[13]. However instead of creating the doublets vocabulary for segmentation and localization, which require too many doublet probabilities to estimate, only the most probable locations of an object within an image are evaluated. Although it was shown that interest point is able to generate nearly accurate hypotheses about localization of objects in the images for a small number of object classes, there is a proportion of visual synonyms and polysemy for a relatively large object classes. 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.

For the object localization and detection system, some studies[14,15] have been conducted to improve both accuracy and speed. In[14], they proposed and evaluated a method that used PCA-SIFT in combination with a clustered voting scheme to achieve detection and localization of multiple objects with reasonable performance. However they typically restricted their method only to two objects. In[15], they demonstrated the feasibility of their approach for relatively large datasets reducing the computational cost. However, the performance of their approach is highly dependent upon the object viewpoints. Moreover, both of the above methods did not perform well for objects with little or no texture. The main limitation of SIFT like feature detectors are that they do not find

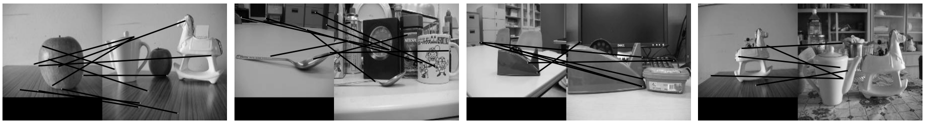


Fig. 1. Some example images from our database on which SIFT detector does not find proper matches

proper matches for such types of objects as illustrated in Fig. 1. Here lines represent matching of SIFT feature points between left and right parts of each of the sub-figure. They perform well only for the feature-rich objects. However, there certainly exist geometrically and texturally simple objects that will not generate many features. Since we have used spatial shape of objects along with color appearance, we have tackled this problem. Combined color and shape information are important aspects in any object recognition system[3]. Representation of shape using the spatial distribution of edges often perform as well as or even better than local patch based detector[11]. On the other hand, color appearance is well-known and powerful information for some objects. Our proposed approach is demonstrated in section 2. A description of datasets of 15 everyday objects in cluttered scenes and different environments is given in section 3. In section 4, we show how our system performs.

2 The Proposed Approach

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, our approach differs in using different features and techniques for both generative and discriminative classifiers. In the generative part, using the pLSA model[18], sets of most probable hypotheses are generated with a bag of visual words. Since the pLSA is mainly used for topic discovery purposes, using it we can easily predict all possible object locations for multiple objects within an image. Moreover, the pLSA shows considerable robustness with respect to partial occlusion, viewpoint and scale changes. On the other hand, the discriminative multi-class SVM classifier is used to verify the sets of hypotheses using both shape and color appearance features. Our integrated approach is comprised of two major steps: *learning the integrated model, and hypothesis generation and SVM verification.*

2.1 Learning the Integrated Model

In our approach, both the pLSA and SVM models are learned to the entire sets of training data using different features sets. Labeled training images containing single and multiple object are used to learn the system. The learning process of the integrated model proceeds in following two parallel stages.

pLSA Model with Bag of Visual Words. In order to generate a bag of visual words for each training and testing images, we first seek visual vocabulary of words that will be insensitive to change in viewpoint, scale and illumination. These visual words are formed by vector quantizing the SIFT descriptors[19] using the K-means clustering. Then the visual vocabulary is obtained by collecting all visual words computed from the training images. The SIFT descriptors are computed on uniformly sampled points detected in object edges over the circular patch with radius $r = 10$. Taking a uniform sample on the edges of an object makes the model shape informative, which is very 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, 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:

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

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 using the Expectation Maximization(EM) algorithm as described in [18] and $P(w, d)$ is given by equation 1.

SVM Model with Merging Features. It has been shown in[16] that pLSA provides a better intermediate representation of images using bag of visual words. On the other hand, object detection algorithms that use discriminative methods combined with global and/or local representations have been shown to perform well in the presence of clutter, viewpoint changes, partial occlusion, and scale variations. Therefore, 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.*[11]. Here the object is represented by its local shape and spatial layout. The local shape is represented by orientations of edge histogram within an object's subregion quantized into K bins 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 subregions 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 (for example, coffee-jar, coke-can). Shape features are not sufficient enough to distinguish among all types of objects (e.g., apple, tape holder). In this case, color appearance is a better feature to find the similarity between them. In order to determine it, we compute 3D HSV global color histograms from all training images collected from different environments in varying lighting conditions. In the HSV color space, quantization of hue requires the most attention. The hue circle consists of the primary colors red, green and blue separated by 120°. A circular quantization at 20° steps sufficiently separate the hues such that the three primaries and yellow, magenta, and cyan are represented each with three sub-divisions. Saturation and value are each quantized to three levels yielding greater perceptual tolerance along these dimensions. Thus H is quantized to 18 levels, and S and V are quantized to 3 levels. The quantized HSV space has $18 \times 3 \times 3 = 162$ histogram bins. The final histogram is normalized to sum to unity in order to fit appropriately in our SVM kernel. The combination of both shape and color appearance features for an image I, are merged as:

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

where both α and β are weights for the shape histogram, $H_S(I)$ and color appearance histogram, $H_A(I)$, respectively. The multi-class SVM classifier is learned using the above merging feature giving the higher weight to the more discriminative feature. The values of α and β in equation 4 are determined for each object separately. We use the LIBSVM package for our experiments in a multi-class mode with the *rbf* exponential kernel.

2.2 Hypothesis Generation and SVM Verification

In the learning stage, both generative and discriminative models are learned on the label training data sets. In this section we have two goals: *the most probable hypotheses generation, and SVM verification and use of context information.*

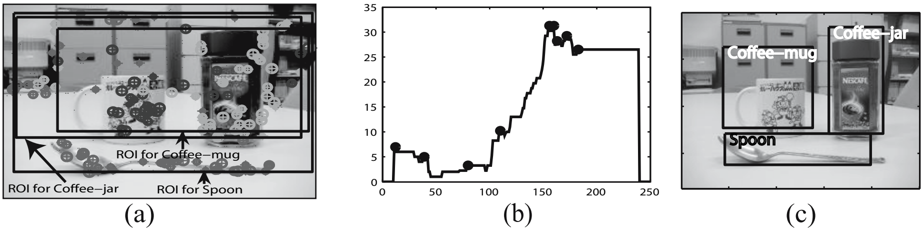


Fig. 2. The most probable hypotheses generation and SVM verification results: (a) three detected ROIs for a new test image (b) local maxima for the object, *coffee-jar* and (c) detected objects for the test image

The Most Probable Hypotheses Generation. In case of pLSA during the learning stage, the model determines the mixture coefficients $P(z_k|d_j)$ for each object d_j (here $z \in z_1, z_2, \ldots z_{15}$ for fifteen objects(topics)). An object d_j is then classified as to maximum $P(z_k|d_j)$ over k. For each new test image all visual words are extracted from all objects and background in the image and each visual word is classified under the topic with the high topic specific probability $P(w_i|z_k)$. Then it is used to detect the region of interest (ROI) for each object instance in the image. The ROI is the smallest rectangular window within the image that contains all possible visual words for a particular object (topic). As an example, Fig. 2(a) shows three ROIs (coffee-jar, coffee-mug and spoon) among possible fifteen ROIs. These ROIs are now used to predict the most probable hypotheses. The pLSA model is mainly used for topics discovery purposes, therefore if there are multiple instances of the same object within an image the model will generate the multiple probable topics representing the same object. The algorithm to find the most probable hypotheses is given below:

1. Find the ROIs for all possible objects within an image based on $P(w_i|z_k)$ and for each ROI repeat the following steps.
2. Compute the average aspect ratio, M_{a_i} of the window for each object 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 training stage using ground truth bounding boxes.
3. Slide the window with the average aspect ratio within each ROI for each object and count the number of visual words for that object within the window.
4. Determine the local maxima based on the average number of visual words in the sliding window Fig. 2(b).
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 most probable hypotheses.

SVM Verification and Context Information. After hypothesis generation, each hypothesis is evaluated using the combined features of both shape and appearance. In the verification step, the features are extracted from the regions of the image bounded by windows of the most probable hypotheses. Therefore, for all windows, shape descriptors and color appearance are combined according to the equation 4 and fed into the multi-class SVM classifier in recognition mode. Only the hypotheses for which a positive confidence measured is returned are kept for each object. Objects with the highest confidence level are detected as the correct objects Fig. 2(c). The confidence level is measured using the probabilistic output of the SVM classifier.

In the post-processing step, the environment related context information is used along with the probabilistic output of the SVM classifier. The context information is determined during the training period from the labeled training images as a co-occurrence table. It is 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 minor increases 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 the office environmental dataset image consists of three objects (book, telephone-set and CD) and one or more of these objects (book, telephone-set) are detected with a high confidence level. Then if the other object(CD) is detected with a low confidence level we include this object in our final detection results. On the other hand, in this case, for objects in other datasets (for example, kitchen environment dataset) a high threshold margin is set to reduce the false positive rate. The base context dataset is determined by using both numbers of detected objects and their probabilities.

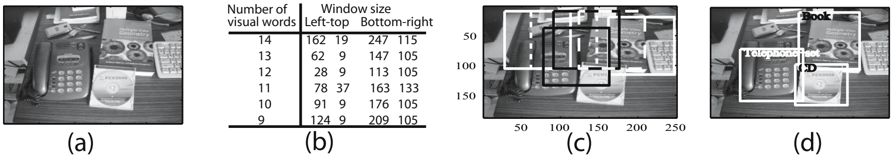
3 Datasets

We are developing a service robot and no standard database is suitable for our application. Therefore, for our experimental purposes, a database is created with ground truth bounding boxes that contain multiple objects per image. It consists of 813 images (1692 objects) of 15 everyday objects related to our application in different environments against cluttered, real-world backgrounds with occlusion, scale, and viewpoint changes. Among these 300 images are single object per image and the rest 513 images (1392 objects) are multiple objects per image. Since objects were presented randomly within an image, therefore, there exist differences in depth, position, rotation and lighting. The depth changes caused a significant amount of scale variation among objects. Fifteen objects were grouped into four datasets. Dataset-1 contains 215 images (80 single and 135 multiple objects per image) of four objects (coffee-jar, coffee-mug, spoon and table-clock). Dataset-2 includes 209 images (80 single and 129 multiple objects per images) of four objects (apple, coke-can, tea-pot and toy-horse). Dataset-3 makes up 150 images (60 single and 90 multiple objects per images) of three objects (book, telephone-set and CD). Dataset-4 consists of 239 images (80 single and 159 multiple objects per images) of four objects (can, stapler, tape-holder and tiffin-box). All images were collected in four different environments: office environment, kitchen environment and two room environments. From 813 images 340 images (300 single object per image and 40 multiple objects per image) were presented during training stage. Another, 473 images were used during recognition step containing multiple objects per image.

4 Experimental Results

In this section we carry out a set of experiments to investigate the benefit of our integrated approach with combined features for multiple object detection and localization on our database. In our experiment, we used 15 different objects collected in different environments and backgrounds. In the training period, both the pLSA and SVM models are fitted for all 15 objects. The object specific feature weights(α and β), optimal threshold value, the penalty parameter(C) and the kernel parameter(γ) are determined using five-fold cross-validation(v=*5*) during the training period of the SVM classifier. 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). In the recognition stage, given an unlabeled image, our main objective is to automatically detect and localize all the objects within an image. The localization performance is measured by comparing the detected window area with the ground truth object window. We count an object as a 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. To understand how the proposed method performs, in the following section we investigate three areas: *benefit of the integrated method, advantages of the merging features and finally, how the dataset related context information improves the overall performance.*

As we previously mentioned , only the generative model is not sufficient enough to detect multiple objects in an image. This is due to the 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, losing all spatial and neighborhood relationship. In our experimental result in Fig. 3, let us consider the original image of Fig. 3(a). In this case, the number of visual words generated for *book* object in different most probable windows is given Fig. 3(b) and their corresponding regions of window is shown in Fig. 3(c). From the illustration it is clear that a significant amount of visual words are generated from the other areas than *book* object due to the visual polysemy nature of objects and/or objects parts and complex backgrounds. However, there is strong evidence among the generated hypotheses for the *book* object in the image and we verified it by our SVM classifier. Fig. 3(d) shows the final detected results by our integrated method for the *book* object along with other two objects(*telephone-set and CD*).



Number of visual words	Window size Left-top		Bottom-right	
14	162	19	247	115
13	62	9	147	105
12	28	9	113	105
11	78	37	163	133
10	91	9	176	105
9	124	9	209	105

Fig. 3. Results of the integrated method: (a)the original test image (b) number of visual words and window size for the object, *book* (c) windows for the object, *book* on the image and (d) detected objects by SVM

The detection performance of our SVM classifier is measured using both shape and color appearance features. In our experiment, the object shape represented by the PHOG and computed within the range 0 to 360° into 40 histogram bins at resolution level $l = 3$. With these parameters selection the second column of Table 1 shows the performance achieved using just PHOGs features. Poor performance is obtained using just PHOGs alone. However, there are some shape informative objects in the datasets for which the detection rate is more than 75%(such as coke-can, coffee-jar etc.). The average localization and detection rate(LDR) for all objects in our database using shape feature alone is 50.25% and average false positive rate(FPR) is 31%. For some objects(e.g.,tape holder, apple, table clock), color information is very relevant and is less confused when using it. Therefore, we use the 3D HSV color histogram with $H = 18$, $S = 3$ and $V = 3$ bins of total 162 dimensional features. The third column of the Table 1 shows the performance of using color features alone. We now, first use the merging features according to equation 4 of both the color and shape appearance with $\alpha = 1$ and $\beta = 1$. The merging features with equal weights give the performance of average LDR is 64.25% and FPR 13.75%. However, the better results are obtained when we use the object-specific weighted merging features as can be shown from the fifth column of the Table 1. The average LDR and FPR are 73.25% and 24.75%, respectively. The weighted merging feature increases the average performance by 10% than the previous one.

Finally, we use the context information as a post-processing stage in our system to improve the overall performance. A fully connected graph between segments label is used to determine the relationship between objects. In our former experiment, in order to reduce false positive rate, the probabilistic output of the SVM classifier for detected objects were compared to the threshold 0.3

Table 1. Experimental results on our datasets

Dataset	Shape feature alone		Color feature alone		Merging feature		Weighted merging feature		Weighted merging feature and context	
	LDR	FPR	LDR	FPR	LDR	FPR	LDR	FPR	LDR	FPR
Dataset-1	0.48	0.34	0.64	0.12	0.68	0.12	0.79	0.21	0.84	0.12
Dataset-2	0.40	0.25	0.40	0.29	0.50	0.13	0.59	0.33	0.65	0.19
Dataset-3	0.65	0.31	0.70	0.22	0.75	0.17	0.84	0.24	0.88	0.11
Dataset-4	0.48	0.34	0.56	0.21	0.64	0.13	0.71	0.21	0.76	0.21

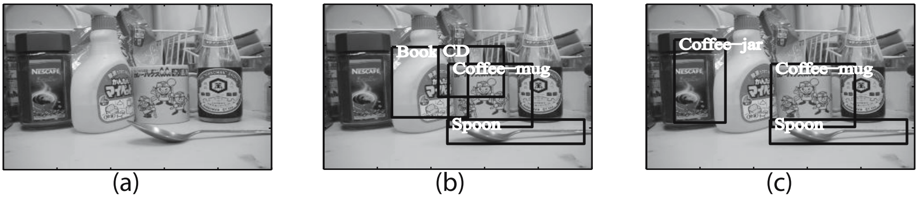


Fig. 4. Result of using context: (a) the original test image with three target objects (*coffee-jar, coffee-mug and spoon*) (b) detected objects without context (c) detected objects with context

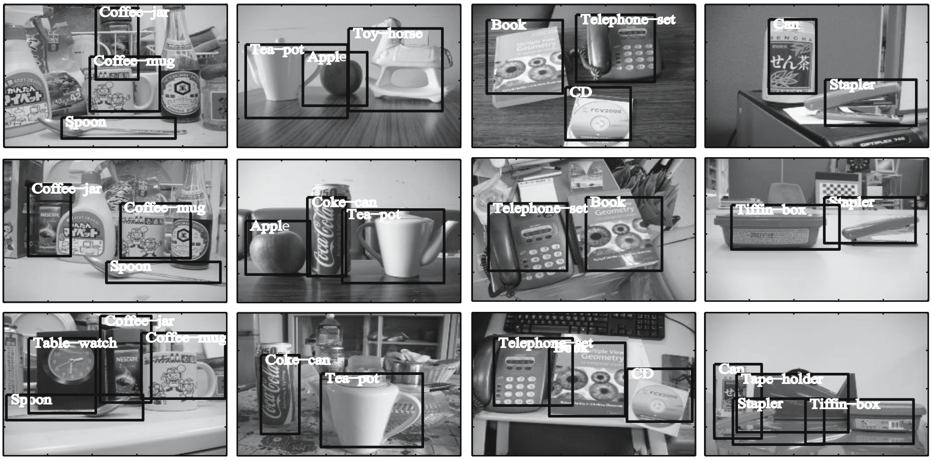


Fig. 5. Detection and localization results on our database

to make the final decision. As a result, some positive objects that are detected with probabilities less than 0.3 are entirely eliminated from our results. On the other hand, some negative objects that are detected with a probabilities greater than or equal to 0.3 are included in our results. In the experimental results for a new test image of Fig. 4(a). The object *coffee-jar* is detected as a *coffee-jar* with the probability 0.28 and not included in the result as shown in Fig. 4(b). The same figure shows that objects *book* and *CD* are included as positive objects and detected with the probabilities 0.33 and 0.37, respectively. Using the context information, determined using the technique as discussed in section 2.2, we set the threshold to 0.5 for non-contextual objects and no restriction is set for context related objects and obtained the final results of Fig. 4(c). The performance on our database using the context information is shown in the last column of the Table 1. We get the best result with average LDR 78.25% and FPR only 15.75% for all fifteen objects.

Our method overcomes the limitations of methods[14,15] and detects all types of objects with a reasonable performance. In[15], texturally simple objects were detected with very poor recognition rate while some texture rich objects detected with 80% of recognition rate. In our experiment, more than 50% of objects are very texturally simple (spoon, apple, stapler etc.) and are detected with 67% of average detection rate. On the other hand, the rest of the texture rich objects are detected with 90% of average recognition rate. Therefore, our method is comparable with the above mentioned methods. Our system also has the ability to detect and localize many objects very fast and can be implemented in real-time. Fig. 5 shows some results of our experiments.

5 Conclusion

We have proposed a new integrated approach for multiple object localization and detection. Our system has shown the ability to accurately detect and localize many objects even in the presence of a cluttered background, substantial occlusion, and significant scale changes. We have demonstrated that the integrated model with merging feature and context information enriches the performance accuracy of the system. In the future we would like to extend this system to detect and localize multiple instances of all types of objects more accurately using more robust features. 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. Furthermore, we will explore the possibility of detecting pose based on the window of the detected object and its surrounding visual words.

Acknowledgements

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

References

1. Seemann, E., Leibe, B., Mikolajczyk, K., Schiele, B.: An evaluation of local shape-based features for pedestrain detection. In: Proc. of British Machine Vision Conference (BMVC 2005), Oxford, UK (2005)
2. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. of European Conference on Computer Vision (ECCV 2004), Workshop on Statistical Learning in Computer Vision, Prague (2004)
3. Diplaros, A., Gevers, T., Patras, I.: Combining color and shape information for illumination-viewpoint invariant object recognition. IEEE Transactions on Image Processing 15, 1–11 (2006)
4. Stella, X., Ralph, G., Jianbo, S.: Concurrent object recognition and segmentation by graph partioning. In: Proc. of Neural Information Processing Systems (NIPS), Vancouver, Canada, pp. 1383–1390 (2002)

5. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Proc. of ECCV 2004, Workshop on Statistical Learning in Computer Vision, Prague, pp. 17–32 (2004)
6. Guillaume, B., Bill, T.: Hierarchical part-based visual object categorization. In: Proc. of International Conference on Computer Vision and Pattern Recognition (CVPR(1)), San Diego, CA, USA, pp. 710–715 (2005)
7. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of model for visual recognition. International Journal of Computer Vision (IJCV) 71, 273–303 (2007)
8. Ferrari, V., Tinne, T., Luc, V.G.: Object detection by contour segmentation networks. In: Proc. of ECCV(3), Graz, Austria, pp. 14–28 (2006)
9. Jacobs, D.: Robust and efficient detection of salient convex groups. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 18, 23–37 (1996)
10. Marcin, M., Cordelia, S.: Spatial weigthing for bag-of-features. In: Proc. of CVPR (2), New York, NY, pp. 2118–2125 (2006)
11. Bosch, A., Zisserman, A., Muñoz, X.: Representing shape with spatial pyramid kernel. In: ACM International Conference on Image and Video Retrieval (CIVR), Amsterdam, The Netherlands, pp. 401–408 (2007)
12. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Group of adjacent contour segment for object detection. PAMI 30, 30–51 (2008)
13. Josef, S., Bryan, R.C., Alexei, A., Zisserman, A., William, T.: Discovering objects and their location in images. In: Proc. of the IEEE International Conference on Computer Vision (ICCV), Beijing, China, pp. 370–377 (2005)
14. Stefan, Z., Manuela, M.: Detection and localization of multiple objects. In: Proc. of Humanoids, Genoa, Italy (2006)
15. Erik, M.C., Jochen, T.: Shared Features for Scalable Appearance-Based Object Recognition. In: Proc. of IEEE Workshop on Application of Computer Vision (WACV), Breckenridge, Colorado, pp. 16–21 (2005)
16. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification using a hybrid generative/discriminative approach. PAMI 30, 712–727 (2008)
17. Fritz, M., Leibe, B., Caputo, B., Schiele, B.: Integrating representative and discriminative models for object category detection. In: Proc. of ICCV, Beijing, China, pp. 1363–1370 (2005)
18. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196 (2001)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110 (2004)