A Hybrid Model for Multiple Object Category Detection and Localization

Dipankar DAS Yoshinori KOBAYASHI Yoshinori KUNO

Department of Information and Computer Sciences
Saitama University, Saitama-shi, 338-8570, Japan
{dipankar, yosinori, kuno}@cv.ics.saitama-u.ac.jp

Abstract

This paper presents a new hybrid approach to simultaneous detection and localization of multiple object categories using both generative and discriminative models. Our approach consists of first learning the generative model (pLSA) and discriminative model (SVM) using bag of visual words and merging features, respectively. Our merging feature combines spatial shape and appearance of an object. At the same time context graphs are generated from the labeled training datasets. Then, given a new unlabeled test image, a set of promising hypotheses are generated for each object category using pLSA model and bag of visual words representing each object. The discriminative part verifies each hypothesis using SVM classifier with merging features. In the post-processing stage, context information along with the probabilistic output of the SVM classifier is used to improve the overall performance of the system. A combination of features and context information are used to investigate the accuracy of the system. 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.

1 Introduction

In the last few years, object detection and localization have become very popular areas of research in computer vision. Although most of the categorization algorithms tend to use local features, there is much more variety on the classification methods. In some cases, the generative models show significant robustness with respect to partial occlusion and viewpoint changes and can tolerate considerable intra-class variation of object appearance[1, 2, 3]. However, if object classes share a high visual similarity then the generative models tend to produce a significant number of false positives. On the other hand, the discriminative models permit us to construct flexible decision boundaries, resulting in classification performance often superior to those obtained by only generative models[4, 5]. However, they contain no localization component and require accurate localization in positions and scale. In the literature, the standard solution to this problem is to perform an exhaustive search over all position and scales. However, this exhaustive search imposes two main constraints. One of them is the detector's computational complexity. It requires large computational time for relatively large number of objects. The second is the detector's discriminance, since a large number of potential false positives need to be excluded.

Our proposed method combines the advantages of discriminative methods with those of probabilistic generative models. The method is based on finding one or more probable locations of an object within an image using a generative model, and then evaluating these locations using a discriminative classifier. In the post-processing stage, the environmental context information is used to improve the overall performance of the system. This paper has three main contributions. The first is a new approach of integrating both generative and discriminative classifiers into a single framework to detect and localize multiple object categories per image. The discriminative part, the SVM verification stage, uses the merging feature of an object to verify these promising hypotheses. The second contribution is that the system automatically generates and uses the context information and the category specific weighted features to improve detection and localization performance. The third contribution is the experimental results show the superiority of the new approach with respect to some state of the art object categorization methods in terms of detection performance and significant reduction of false positive rate.

It has been recently shown that combining the power of generative modeling with a discriminative classifier allows us to obtain good localization and categorization[6, 7]. However, the proposed hybrid approach in [6] was mainly used for scene classification and did not provide any location information of the object. On the other hand, in [7] 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. In our previous research, we used a combination of pLSA and discriminative classifier for detection and localization of specific object. However, our approach in this research differs from the previous one with respect to the following: (i) our discriminative classifier uses more reliable shape and appearance features to detect and localize large number of object categories (ii) more efficient algorithm is designed and implemented to generate promising hypotheses, and finally (iii) we do experiments over some standard datasets to compare the performance of our method with some state of the art recognition frameworks.

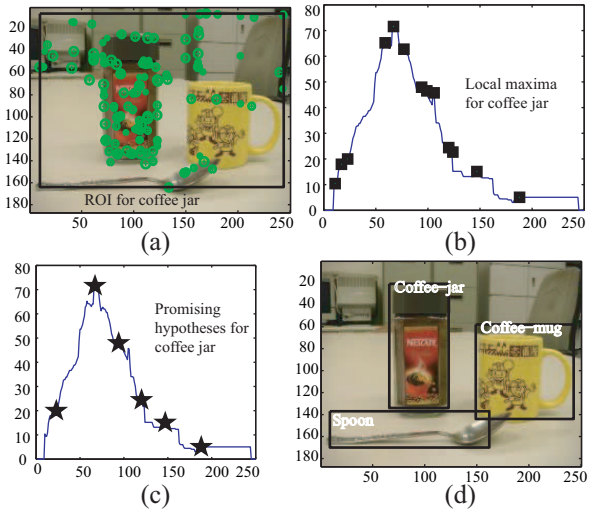

Figure 1: Hypotheses generation and SVM verification

2 Learning the Generative Model

To fit the pLSA model[8], we first seek vocabulary of visual words for training images that will be insensitive to change in viewpoint, scale, and illumination. This vocabulary is formed by vector quantizing the SIFT descriptors[9] using the K-means clustering algorithm. The SIFT descriptors are computed on uniformly sampled points in object edges over the circular patch with radius $r = 10$. After constructing the visual vocabulary, in the formulation of pLSA for images[1], a co-occurrence table is computed where each image is represented as a collection of visual words. The pLSA model associates each observation of a visual word, w within an image, d with a topic variable, $z \in Z\{z_1, z_2, \ldots, z_k\}$. Here our goal is to determine $P(w|z)$ and $P(z|d)$ by using the maximum likelihood principle. The model is fitted for all training images using the Expectation Maximization(EM) algorithm [8] to a number of topics k. 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 belongs to multiple sub-topics and our model automatically fits to an optimal number of sub-topics.

3 Promising Hypotheses Generation

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. For simplicity, among three detected ROIs, the Fig. 1(a) 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.
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.
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.
4. Determine the local maxima (Fig. 1(b)) 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, which overlap by 75% or more with the window that contains the maximum number of visual words for each local region.
6. After suppressing the non-maximum windows in each neighborhood the remaining windows are selected as the promising hypotheses (Fig. 1(c)).

4 SVM Learning and Verification

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.*[10]. Our final shape descriptors represented by the normalized Pyramid Histogram of Orientation Gradient(PHOG) and computed within the range 0 to 360° into 40 histogram bins at resolution level $l = 3$. 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 section 2. Then the normalized histogram of visual words for each object is computed. Finally, the combination of both shape and appearance features for an object O, are merged as:

$$H(O) = \alpha H_S(O) + \beta H_A(O) \qquad (1)$$

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 equation 1 are determined for each object separately. We use the LIBSVM[11] 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 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. 1(d)). The confidence level is measured using the probabilistic output of the SVM classifier.

5 Post-Processing using Context

In the task of object category recognition, environmental context information can play an important role of reducing ambiguity in an object's visual appearance. To incorporate context 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. 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 selected. The base context information is then used to give the flexible margin for the context-related objects and hard margin for non-contextual objects. It is mainly used to improve the detected performance of the SVM classifier.

6 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 experimental results, 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.

Comparison with Other Methods. For comparison purposes, most of the datasets are collected from PASCAL VOC database collection. The performance of our system is compared to the integrated representative and discriminative (IRD) representation of Fritz *et al.*[7], the implicit shape model (ISM) of Leibe *et al.*[3] and local kernels (LK) representation of Wallraven *et al.*[12], using the same datasets that are tested in[7]. For each dataset we use the SVM classifier with PHOG feature to verify the hypotheses generated by our algorithm as discussed in section 3. Table 1 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.

Table 1: Performance comparison with other methods

Category and Dataset	LK[12]	ISM[3]	IRD[7]	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%

Table 2: Hypotheses generation and verification results

Category	Detected objects							Undetected object	SVM 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**

Each image in these datasets contains only one target object. 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.

Benefits of the Integrated Method. One of the main contributions of this paper is to integrate both generative and discriminative approaches into a consistent framework. 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. In this experiments, we use our own 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. Table 2 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 window 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 2 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 2. In this section, we also investigate how our method performs on MIT-CSAIL static office datasets for three categories of objects: computer monitors, computer keyboards and bookshelves. Our final result is comparable with Sivic *et al.*[1] 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. However, in our approach the detection and localization accuracy for computer screen and bookshelf are 84% and 93%, respectively. We obtained recognition accuracy of 77% for computer keyboard. Our better performance compared to[1] could be due to the integration of both generative and discriminative classifiers instead of using only generative model.

Table 3: Experimental results on authors' datasets

Category	Merging feature (MF)		MF with context		Weighted MF (WMF)		WMF 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**

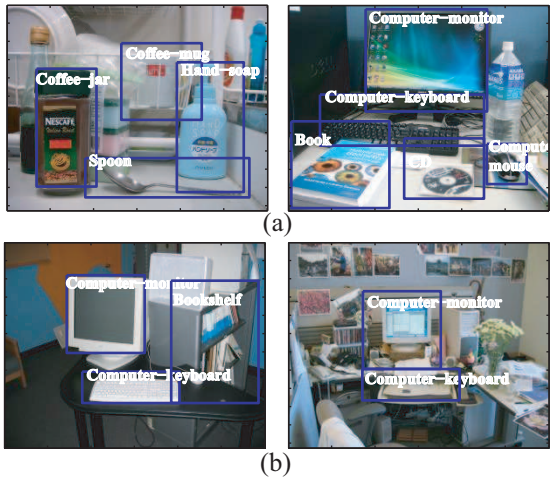

Figure 2: Example detection and localization results: (a) authors' datasets (b) MIT-CSAIL datasets

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[1, 7, 3], 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. 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. Table 3 shows the detection and localization rate (DLR) at the false positive rate (FPR) indicated in their adjacent column. The merging feature without any weight and context information produces an average DLR 66%. However, when the same feature is used 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 and localization results on our own and MIT-CSAIL datasets are shown in Fig. 2.

7 Conclusion

In this research, 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. Our experimental results demonstrated 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 feature and category specific weighted merging feature 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. In the future, we will explore the possibility of detecting pose based on the window of the detected object by SVM classifier and its surrounding 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 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] S. Josef, C. Bryan, Russell, A. Alexei, A. Zisserman, and T. William, "Discovering objects and their location in images," *Proc. ICCV'05, Beijing, China*, pp.370–377, 2005.

[2] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of model for visual recognition," *IJCV*, vol.71, no.3, pp.273–303, 2007.

[3] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," *Proc. ECCV'04, Prague*, pp.17–32, 2004.

[4] A.Y. Ng, and M.I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes," *Proc. NIPS'01, Vancouver, British Columbia, Canada*, pp.841–848, 2001.

[5] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Group of adjacent contour segment for object detection," *PAMI*, vol.30, no.1, pp.30–51, 2008.

[6] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *PAMI*, vol.30, no.4, pp.712–727, 2008.

[7] 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.

[8] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol.42, no.1/2, pp.177–196, 2001.

[9] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol.60, no.2, pp.91–110, 2004.

[10] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with spatial pyramid kernel," *Proc. CIVR, Amsterdam, The Netherlands*, pp.401–408, 2007.

[11] C.C. Chang and C.J. Lin, "Libsvm: A library for support vector machines.", 2008.

[12] C. Wallraven, B. Caputo, and A.B.A. Graf, "Recognition with local features: the kernel recipe," *Proc. ICCV*, pp.257–264, 2003.