Recent Progress on Object Classification and Detection

Tieniu Tan, Yongzhen Huang, and Junge Zhang

Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences, (CASIA), Beijing, China
{tnt,yzhuang,jgzhang}@nlpr.ia.ac.cn

Abstract. Object classification and detection are two fundamental problems in computer vision and pattern recognition. In this paper, we discuss these two research topics, including their backgrounds, challenges, recent progress and our solutions which achieve excellent performance in PASCAL VOC competitions on object classification and detection. Moreover, potential directions are outlined for future research.

Keywords: Object classification, Object detection, PASCAL VOC.

1 Introduction

Object classification and detection are two core problems in computer vision and pattern recognition. They play fundamental and crucial roles in many applications, e.g., intelligent visual surveillance, image and video retrieval and web content analysis. Object classification and detection share some common components and face many common challenges (see Fig. 1), e.g., variability in illumination, rotation and scales, as well as deformation, clutter, occlusion, multi-stability and large intra-class variations.

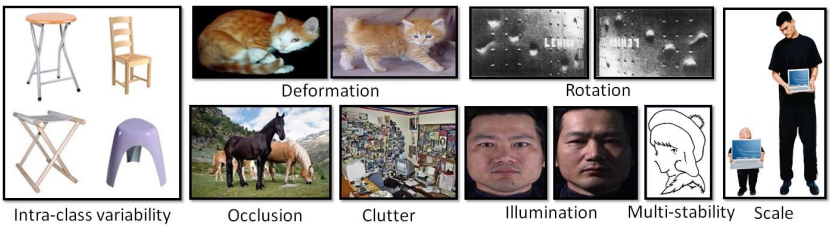

Fig. 1. Common challenges in object classification and detection

Despite the above challenges, great progress has been made in the past decades, and many algorithms have been proposed. In this paper we discuss the general framework of object classification and detection, and some classic methods in each

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 1–8, 2013.

component of the framework. Afterwards, we introduce our work on object classification and detection, especially our solutions which were ranked among the best in the PASCAL VOC competition [1,2]. Finally, we point out some potential directions for future work.

2 General Framework and Methods

2.1 General Framework

The general framework of object classification and detection is illustrated in Fig. 2, where the first and the last module are shared by object classification and detection. In this subsection, we discuss these two common modules, and then analyze the other modules in the following two subsections.

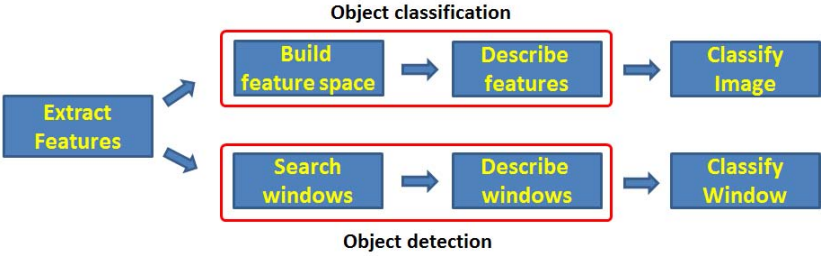

Fig. 2. A general framework of object classification and detection

The first module, i.e., feature extraction, usually includes two main steps: extracting image patches and representing image patches. Extracting image patches is implemented via sampling local areas of images, usually in a dense or a sparse manner. Representing image patches is implemented via statistical analysis over pixels of image patches. The representation vectors of image patches are called local features. Widely used features include: 1) appearance based ones, e.g., scale-invariant feature transform (SIFT) [19], histogram of oriented gradients (HOG) [8]; 2) color based ones, e.g., color descriptors [25]; and 3) texture based ones, e.g., local binary pattern [22] and Gabor filter [18].

The last module, i.e., classification, is a hot topic in machine learning. Many classic classifiers are used in object classification and detection, e.g., Boosting, SVM and KNN. Also, kernel tricks, e.g., inter section kernel [3], are often used to enhance the overall performance.

2.2 Object Classification

In this subsection, we discuss the other modules in object classification, i.e., building feature space and describing features with the feature space.

Building Feature Space. Feature space consists of a group of base vectors. In particular, in the well-known bag-of-features model, the feature space is composed of a set of dictionaries, which are also called visual codes or codebook. There are three strategies to build the feature space, explained as follows.

The first one randomly chooses patches from images as the base vectors. This method is adopted in some biologically inspired models [27,14,13]. It is fast but does not sufficiently reflect the characteristics of the feature space.

The second one is based on supervised learning, i.e., generating dictionaries via supervised learning over local features. This scheme builds the relation between features and image labels, and well reflects the structure of the feature space. However, it is time-consuming because it needs to iteratively resolve dictionaries. For more details, readers are referred to the literature [5,20].

The third one is based on unsupervised learning, i.e., obtaining the base vectors via unsupervised learning over local features. This strategy finds a good balance between accuracy and speed, and is widely used in recent methods.

Describing Features. Describing features is a very important component in object classification, and greatly influences image classification in both accuracy and speed. Existing coding strategies can be divided into the following categories:

Voting-based methods [7,12] describe the distribution of local features, reflecting the occurrence information of visual codes.

Fisher coding-based methods [23,24] calculate the distribution of local features with the Gaussian Mixture Models. Each Gaussian model describes a kind of local features.

Reconstruction-based methods [32,29] encode a feature by least-square-based optimization with constraints on the number of codewords for reconstruction.

Local tangent-based methods [33,38] firstly estimate the manifold of the feature space, based on which an exact description of local features is derived.

Saliency-based methods [15,31] depict a local feature by the saliency degree, e.g., the ratio of the distances from a local feature to the codewords around it.

For more details of feature coding, readers are referred to our recent paper [16], which provides a comprehensive study about feature coding.

2.3 Object Detection

A typical object detection system is composed of four major steps: window search, object representation, machine learning and optimization, and post-processing. For the sake of space, we only introduce window search and object representation in this subsection.

Window Search: Most existing approaches follow the sliding-window paradigm [8,10,28,35,26,9]. In this case, generic object detection is formulated as a binary classification task. In the detection stage, the detector model evaluates each sliding window across scales and locations in an image, and then thresholds it as an object or not. So the simplest method is applying exhaustive search. In

contrast to exhaustive search, there are several approaches on heuristic window search for the purpose of narrowing search space and accelerating detection.

Exhaustive search can be found in typical detection methods [8,35,9]. The advantage of exhaustive search is that it can obtain a relative high recall rate, reducing missing rate. But the large search space makes it intractable for real time object detection. Motivated by this challenge, heuristic window search methods have been developed. Lampert *et al.,*[17] propose efficient subwindow search (ESS) with branch and bound strategy. This method is based on sparsely detected features. If it is built on densely extracted features, the efficiency of ESS cannot be guaranteed. We all know that the segmentation and salience information provide the prior for object location. Thus, there are also some studies on heuristic window search based on segmentation and salience analysis [28].

Object Representation: Representation models mainly include part based models [35,36,26,9,34] and rigid template models [8,28].

Part based representation was firstly proposed by Fischler and Eschlager [11]. In this model, an object is represented by several parts in a flexible and deformable configuration [11]. Each part is usually described or represented by a small rigid template. The spatial relationships between parts are considered by spring-like connections for structure description [6]. Therefore, part based model can be considered as a top-down structured model, which is robust to partial occlusion and appearance variations. Recently, excellent part based models [26,9] have shown their success on many difficult datasets [21]. Due to their robustness to deformation, occlusion, part based model is regarded as a promising method for object localization.

The other common representation model is the rigid template model [8,35]. Rigid template model describes an object in a holistic manner, so they cannot well capture the structure variations of objects. Therefore, they perform well on well conditioned databases but suffer from those challenging data with deformations and occlusions. Besides, both the part based model and the rigid template model are associated with low-level features. Thus, progress on low-level features helps the improvement of object representation greatly as well. One classic feature is histogram of oriented gradients (HOG) [8]. The others include scale-invariant feature transform (SIFT) [19] based bag-of-words feature, pairs of adjacent segments (PAS) [10], local binary patterns (LBP) [22],*etc.*

3 Our Work

In this section, we introduce our solutions to object classification and detection, which were ranked among the best in the PASCAL VOC competition.

3.1 Object Classification

Our system of object classification is based on the bag-of-features model [7], with four steps different from traditional ones as illustrated in Fig. 3 and explained as follows.

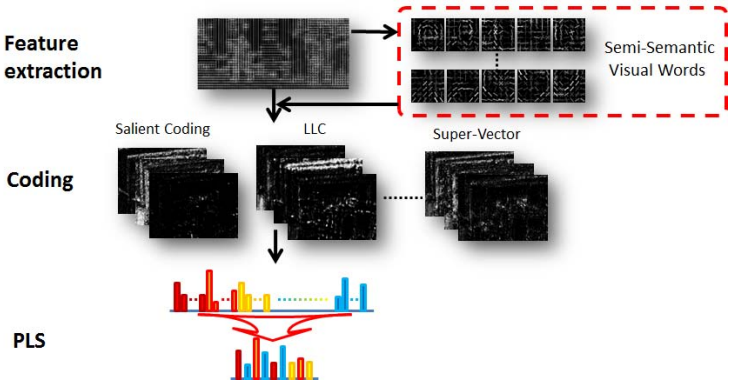

Fig. 3. Our system of object classification in VOC

1. We use different kinds of low level features: SIFT, HOG, LBP, color descriptor and Gabor filters. They play different roles in object description.
2. We apply parts learning from the deformable part-based model [9]. The learnt parts are augmented as a new set of dictionaries, which is demonstrated to be discriminative in differentiating objects from different classes.
3. We choose three methods for feature coding: super-vector coding [38], local constraint liner coding [29] and salient coding [15]. These methods complement each other in feature coding, which helps to improve the overall performance of object classification.
4. With different feature description and feature coding methods, the final dimensionality is very high. For dimensionality of more than a million for each image, which dimensionality reduction method should we use? We find that PLS [30] is good choice in this case.

3.2 Object Detection

Our detection system is based on the local structured model. In VOC2010, at the feature level, we propose Local Structured Descriptor and develop new descriptors from shape and texture information, respectively. Secondly, at the topology level, we present a local structured part representation with boosted feature selection and fusion scheme. Fig. 4(a) shows the framework we used in VOC2010.

The system includes two parts: building features and training local structured part detectors. The first part includes extracting local structured descriptors and feature selection of local structured LBP in a supervised manner. In the second stage, we firstly train the root model or holistic model using the learnt features from the first stage, then initialize local structured part appearance models from the root model. Linear SVM is applied to optimize the parameters. For more details, we refer readers to our previous papers [35,34].

The method mentioned above has two limitations: the model complexity is high and the model is still not “deformable” enough. Motivated by these challenges, we propose a data decomposition and spatial mixture modeling method

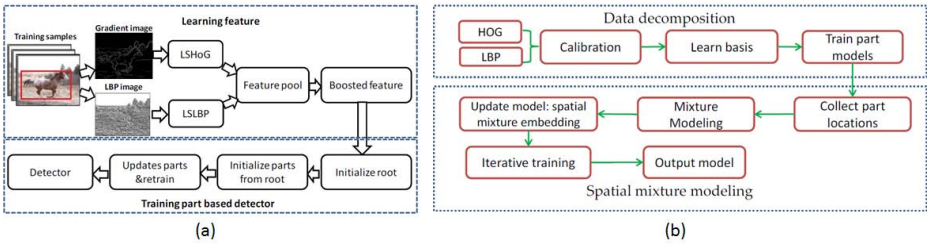

Fig. 4. System framework used in VOC2010 [35] and VOC2011 [36]

[36,37] in VOC2011 as shown in Fig. 4(b). Firstly, data decomposition is developed for the part based model, which not only largely reduces memory usage and computational cost but also outperforms other related systems. Secondly, a spatial mixture modeling method in which part location is described as a mixture distribution learnt from weakly labeled data, is proposed for more flexible structure description. Thirdly, we unify the spatial mixture model into the data decomposition framework. To the best of our knowledge, the presented system achieves the state-of-the-art performance compared with all other related methods from both the competition and the open literature. Due to limit of space, we refer readers to [36,37] for details.

4 Future Work and Conclusions

Object classification and detection, despite of decades research, remain two very active research topics in computer vision and pattern recognition. Every year, more than one hundred related papers appear in various top conferences and journals. Also, it should be recognized that there are still many challenges to be solved as we discussed in Introduction. Based on the analysis in this paper and our own experience, we think that the following directions deserve more attention:

- For object classification, the spatial relations of local features and the structure information of objects are still not well exploited. We believe that the progress in these two problems will greatly enhance current object classification models. Besides, representation learning [4] has shown good potential and provide new perspective on understanding object classification.
- For object detection, current methods mainly focus on learning structure parameters from data but ignore jointly learning structure topology and structure parameters. A potential direction is learning them together from data. Secondly, big data not only bring challenges but also opportunities for object detection. How to take advantage of big data with the latest machine learning techniques and biological observations is also promising.
- In addition, the integration of object classification and detection at the representation level is also an interesting direction for future work.

Acknowledgments. This work is jointly supported by National Basic Research Program of China (2012CB316300) and National Natural Science Foundation of China (61135002, 61203252).

References

1. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/index.html
2. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/index.html
3. Barla, A., Odone, F., Verri, A.: Histogram intersection kernel for image classification. In: Proc. IEEE Inter. Conf. Image. Process. (2003)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE T-PAMI 35(8), 1798–1828 (2013)
5. Bradley, D.M., Bagnell, J.A.: Differential sparse coding. In: Proc. Neu. Inf. Process. Sys. (2008)
6. Crandall, D., Felzenszwalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2005)
7. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Proc. Eur. Conf. Comput. Vis. (2004)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2005)
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE T-PAMI 32(9), 1627–1645 (2010)
10. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of adjacent contour segments for object detection. IEEE T-PAMI 30(1), 36–51 (2008)
11. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Trans. Comput. C-22(1), 67–92 (1973)
12. van Gemert, J.C., Geusebroek, J.-M., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 696–709. Springer, Heidelberg (2008)
13. Huang, Y., Huang, K., Tao, D., Tan, T., Li, X.: Enhanced biological inspired model for object recognition. IEEE T-SMC-Part B 41(6), 1668–1680 (2011)
14. Huang, Y., Huang, K., Tao, D., Wang, L., Tan, T., Li, X.: Enhanced biological inspired model. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2008)
15. Huang, Y., Huang, K., Yu, Y., Tan, T.: Salient coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2011)
16. Huang, Y., Wu, Z., Wang, L., Tan, T.: Feature coding in image classification: A comprehensive study. IEEE T-PAMI (accepted, 2013)
17. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Beyond sliding windows: Object localization by efficient subwindow search. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2008)
18. Lee, T.S.: Image representation using 2D gabor wavelets. IEEE T-PAMI 18, 959–971 (1996)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110 (2004)
20. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Supervised dictionary learning. In: NIPS (2008)
21. Mark, E., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV 88(2), 303–338 (2010)

22. Ojala, T., Petikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proc. IAPR Inter. Conf. Pattern Recognit. (1994)
23. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2007)
24. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
25. Sande, K., Gevers, T., Snoek, C.: Evaluation of color descriptors for object and scene recognition. IEEE T-PAMI 32(9), 1582–1596 (1998)
26. Schnitzspan, P., Roth, S., Schiele, B.: Automatic discovery of meaningful object parts with latent crfs. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2010)
27. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE T-TPAMI 29(3), 411–426 (2007)
28. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proc. IEEE Inter. Conf. Comput. Vis. (2009)
29. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2010)
30. Wold, H.: Partial least squares. In: Encyclopedia of Statistical Sciences (2004)
31. Wu, Z., Huang, Y., Wang, L., Tan, T.: Group encoding of local features in image classification. In: Proc. IAPR Inter. Conf. Pattern Recognit. (2012)
32. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2009)
33. Yu, K., Zhang, T.: Improved local coordinate coding using local tangents. In: Proc. Int. Conf. Mach. Learning (2010)
34. Yu, Y., Zhang, J., Huang, Y., Zheng, S., Ren, W., Wang, C., Huang, K., Tan, T.: Object detection by context and boosted hog-lbp. In: ECCV workshop on PASCAL VOC (2010)
35. Zhang, J., Huang, K., Yu, Y., Tan, T.: Boosted Local Structured HOG-LBP for Object Localization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2011)
36. Zhang, J., Huang, Y., Huang, K., Wu, Z., Tan, T.: Data decomposition and spatial mixture modeling for part based model. In: Proc. Asi. Conf. Compt. Vis. (2013)
37. Zhang, J., Yu, Y., Huang, Y., Wang, C., Ren, W., Wu, J., Huang, K., Tan, T.: Object detection based on data decomposition, spatial mixture modeling and context. In: International Conference on Computer Vision Workshop on Visual Object Classes Challenge (2011)
38. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 141–154. Springer, Heidelberg (2010)