Murat Ambarkutuk {murata@vt.edu}, 1/26/2016

1 A discriminatively trained, multiscale, deformable part model

In their paper, the authors claim that even though the "bag-of-X" representation for object detection is conceptually weaker, it sometime is more successful than more intutive approaches. The authors propose a technique for object detection problems, in which objects are defined with their parts and detected in a multiscale manner. The problem is that being addressed in the paper is to model the objects with their parts, an abstraction that it is difficult to model with BoW approaches. The intuitive explanation of the proposed method is to detect the rough position of an object in a coarsely sampled-multiscale descriptor space. The proceeding step then is to localize the parts of the of the objects in finely defined descriptor space and try to hypothesize the location of the parts. The efficacy of the technique is analizyed in VOC 2007 dataset and the results are compared with other proposed approaches dealing with the same problem.

2 Approach

The proposed approach combines a robust human detection technique (Dalal and Trigss, 2005) with the part-based object modeling framework (Felzenzwalb and Huttenlocher, 2000) to address the category level object detection problem. The former technique describes the images with histogram of gradient (HOG) while the latter formulates the objects with their parts. The proposed method first discretizes the images with 8-by-8 image regions; within that regions HOG's are accumulated. By doing so, the method provides a locally rotation invarient frame work. The next step is to localize the object in that roughly defined description space which forms the initial hypothesis about the objects position. Also these steps are repeated at different scales to create an image pyramid which makes the proposed method scale invarient. The successive step is to minimize deformation costs which results from placing the parts of the object within that region relative to the "root" of the object localized in earlier step. Combining these two methods images can be considered as the first contribution of the paper. The second contribution of the paper is the generalized formulation of SVM's which can handle latent variables, such as the position of the parts in the context of the paper. The proposed learning framework provides a new training framework based on "hard negative" examples, which enables the method to learn from a subset of the whole dataset, which is an important subject if the dataset to be learnt is big.

3 Conclusions

VOC 2007 dataset was used to assess the overall performance of the proposed method. The method outperfomed in detecting objects in more than half of the categories when it is compared with the other methods. In their tutorial session, the authors mention that the HOGs method proposed by (Dalal and Trigss, 2005) showed %79 percent average precision in one dataset while %12 in other. This discrepancy shows that the category level object detection is still an open question. One of the reason why the proposed method outperformed in many object classes that the combination of the flexibility provided by (Felzenzwalb and Huttenlocher, 2000) and robustness of (Dalal and Trigss, 2005). However, one question that comes to one's mind is that how the algorithm would response if there are missing part in the object being detected; for instance, a common case for viewpoint change. It would be a more interesting paper to read if they provide some details regarding that issue.