Commentary: Unsupervised Category Modeling, Recognition, and Segmentation in Images

Anant Krishna Mahale
z5277610

1. Introduction

The reviewed paper addresses three different computer vision problems in real world which include detection, recognition and segmentation of objects by learning a structural model of the category in terms of photometric, geometric, and topological properties of 2D images in unsupervised fashion.

The author aims eliminates the painstaking job of specifying the image category manually by a human being by defining a group of sub-images that share common geometric, photometric and topological characteristics. A tree that captures a multiscale segmentation of the image is represented by each image. To extract the maximally matching subtrees throughout the set, the trees are paired, which are taken as instances of the target category. The extracted subtrees are then combined into a tree union that represents the canonical category model. By finding matches of the category model with the segmentation tree of a new image, detection, recognition, and segmentation of objects from the learned category are achieved simultaneously. This approach is not affected by difference in orientations, scale, occlusion, small appearance changes or noise, region shape deformations and clutter of the test images.

The category modeling, recognition and segmentation are active research topics and have their own applications in real life. These can potentially solve object detection and segmentations in various fields like photography, autonomous car driving, plant phenotyping, in medical industry for detecting and tracking of the cells as it automatically categorize the label in unsupervised fashion.

2. Methods

Three proposed method comprises of four parts: selection of region properties and construction of segmentation tree, tree matching algorithm, learning the model based on the category and optimal weighting of region properties.

For the selection of region properties and construction of segmentation tree, a multiscale segmentation algorithm is used. It partitions an image into homogeneous regions of a priori unknown shape, size, gray-level contrast, and topological context at different range intensity contrasts. Thus, all the segmented regions which represents minute details of the image, constitute child of the tree and the whole image represents the root of the tree. The number of nodes, branching factor and number of levels in different part of the tree depended on the image and it varies.

The authors consider not only intrinsic photometric and geometric properties but also relative interregional characteristics that describe the region's spatial layout and its neighbors. In this way they can capture each and every minute details of the picture. Each of the node in the tree is denoted by v defined by a vector of the respective region's properties Ψ_v. Finally, Ψ_v is normalized over multiscale regions of the training images. However, the author does not take consideration of deep-learning for multi-instance image segmentation since deep-learning methods outperforms traditional computer vision techniques and can be scaled and generalized for different datasets.

For the tree matching method, all pairs of segmentation trees are matched to identify those pairs with a similarity measure above a selected threshold to extract recurring comparable subimages from the given image set which incorporates mutual containment of regions resulting improved robustness of cross-image region matching. The authors also, account for the fact that some of the details such as low contrast in some of the images might preserved less in some images which intern changes the segmentation tree structure by using well known well-known framework of edit-distance graph matching technique [1], [2], [3]. The approach presented in this paper is an extension of Torsello and Hancock's approach [2] by searching for correspondence between regions as well as between groups of contiguous regions in two segmentation trees.

The authors also compare various approaches to edit-distance matching as edit-distance methods are on the assumption one to one node correspondence [1], [2]. To address this issue, the authors considers many to many matching [4] using a technique where the subset of the graph nodes is combined into a single node (merger) when the gap between their attributes is less than the threshold chosen. However, the downside of this merger is that it is heavily influenced by the chosen threshold. Even though the paper [3] talks about many to many relations, it does not account for bias favoring one-to-one node correspondences over one-to-many.

Based on the methods stated above, the selection of threshold influences the structure of the tree, this paper does not address the selection of right threshold. The authors use a straightforward strategy based on the frequency histogram of all the nodes in the tree. Though this method might work for a particular dataset, it might not work in real life applications, since the histogram frequency depends on various factors.

Outliers are accounted for by the author when creating various tree unions. Since most trees are likely to represent instances of categories, a small entropy will characterize node frequencies. The authors therefore get a set of tree unions for DAG permutations in order to learn the category model. Then, for each node the frequency of its matches with nodes in DAG is calculated. The best tree union approximation is chosen based on entropy that achieves a minimum for the sets containing all isomorphic trees. Therefore, the permutation for which entropy is lowest for all permutations is chosen to measure tree node frequency as the best tree union approximation. The tree union with the least number of nodes is used in the case of multiple solutions.

The empirical distribution of node-property variations is computed for each node pair. The histogram obtained is modelled as a Gaussian mixture density of two elements. The mean, Gaussian distribution variances and mixing coefficients are determined using the EM algorithm[5]. All node pairs are split into two mutually exclusive subsets, 1 and 2, which correspond to two Gaussian mixture density components.

3. Results

The authors have used qualitative evaluation for the tree union models learned on the arbitrary image set and quantitative evaluation methods for the detection, recognition, and segmentation of all instances of a learned category present in a test image. To perform the evaluation, the authors have used four different datasets Caltech-101[6], Caltech rear-view cars [7], UIUC multiscale side-view cars and Weizmann side-view horses [8] and TUD side-view cows [9].

The authors have considered various evaluation metrics since they address 3 different problems addressed in the reference paper. They carry out Object detection, recognition, and segmentation jointly by matching the learned tree union model with the image-test trees. The common subtrees of whose measure of similarity is greater than the specified threshold is adjusted as detected objects. The authors prefer RPC (recall-precision curve) as a preferred measure of performance with respect to object detection and segmentation compared to those used by classification-based techniques.

Authors have carried out 3 different experiments with different setting to evaluate the performance of the proposed methods. Object detection and segmentation are relatively accurate for small training dataset in spite of slight image rotations, blur, and partially occluded. However, there are many instances where image is not detected.

One of the flaws in the method can be seen, when model is tested against car images. Since car windows reflect the surrounding regions which vary from image to image, the windows images in the test images, won't be part of learnt model. Thus, it is difficult for the model to generalize the objects when tested against the unseen images.

From the experimental results, it is clear that number of positive training images have an impact on the performance of the model. As the number of positive images increases, the F-measure increases.

Today, Dice Similarity Co-efficient (DSC) and Intersection over Union (IOU) [10] measures are widely accepted measures for image segmentation. Considering this fact the measured employed by the author cannot be used for bench marking the results.

4. Conclusions

The reviewed paper, presents an unsupervised learning approach to categorize, recognize and segment images. It relies on geometric, photometric, and topological properties of the given dataset. The authors have suggested using a many-to-many matching algorithm that identifies matching subimages within each pair of images to identify category occurrences in the unlabeled image set. In terms of differences in geometric, photometric, and topological properties of subregions embedded within the subimages, authors have specified a new measure of similarity between matching subimages that is recursively computed. This test of similarity fuses details on the similarities of the embedded subimages, where the similarities are weighted in proportion to their relative identification importance. Without using any supervision techniques, the authors have presented an algorithm for estimating these weights. The results are also compared to the baseline methods in every category.

Even though the authors were successful in getting near benchmark results, they fail to address many real-world problems. Any unsupervised learning method should be able to handle amount of positive and negative images. In other words, there should not an impact on the performance of the model based on the positive and negative set of images as the whole purpose of performing unsupervised method is to categorize/label the data. The authors use only small dataset ~60 images to evaluate the performance which is very small in todays world. The fact that the structure of the tree depends on the threshold which intern influences the performance is not a good choice since the model can perform poorly with bad choice of the threshold value.

Today many deep-learning algorithms[11] have outperformed traditional computer vision techniques not only in terms of performance but also in terms of generalization for the unseen images which is more practical and useful while building the applications.

References

[1]H. Bunke and G. Allermann, “Inexact Graph Matching for Structural Pattern Recognition,” Pattern Recognition Letters, vol. 1, no. 4, pp. 245-253, 1983.

[2]A. Torsello and E.R. Hancock, “Computing Approximate Tree Edit Distance Using Relaxation Labeling,” Pattern Recognition Letters, vol. 24, no. 8, pp. 1089-1097, 2003.

[3] T.B. Sebastian, P.N. Klein, and B.B. Kimia, “Recognition of Shapes by Editing Their Shock Graphs,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 5, pp. 550-571, May 2004.

[4] M. Pelillo, K. Siddiqi, and S.W. Zucker, “Many-to-Many Matching of Attributed Trees Using Association Graphs and Game Dynamics,” Proc. Int’l Workshop Visual Form, vol. 2059, pp. 583- 593, 2001.

[5] A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. Royal Statistical Soc. B,
vol. 39, no. 1, pp. 1-38, 1977.

[6] L. Fei-Fei, R. Fergus, and P. Perona, “One-Shot Learning of Object Categories,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 594-611, Apr. 2006.

[7] R. Fergus, P. Perona, and A. Zisserman, “Object Class Recognition by Unsupervised Scale-Invariant Learning,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 264-271, 2003.

[8] E. Borenstein and S. Ullman, “Class-Specific, Top-Down Segmen-tation,” Proc. European Conf. Computer Vision, vol. 2, pp. 109-124,2002.

[9] B. Leibe, A. Leonardis, and B. Schiele, “Combined Object Categorization and Segmentation with an Implicit Shape Model,” Proc. ECCV Workshop Statistical Learning in Computer Vision, pp. 17- 32, 2004.

[10] Wang, Z., Wang, E. & Zhu, Y. Image segmentation evaluation: a survey of methods. Artif Intell Rev 53, 5637–5674 (2020). https://doi.org/10.1007/s10462-020-09830-9

[11] X. Ji, A. Vedaldi and J. Henriques, "Invariant Information Clustering for Unsupervised Image Classification and Segmentation," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 9864-9873, doi: 10.1109/ICCV.2019.00996.