Murat Ambarkutuk {murata@vt.edu}, 03/15/2016

1 Names and Faces in the News

In their paper, the authors propose a method to localize objects in images by leveraging keyword lists corresponding to images. The most significant contribution of the paper is a learning model infering positions and scales of objects given the tags corresponding to the image. The general idea of the contribution is to capture the implicit reasons why specific words are choosen in the specific order by the annotater.

2 Approach

The authors assert that given an image with corresponding keywords, the order of the keywords captures implicit properties of the image. As we discussed in the last paper, visual attention provides strong information if it is used in detection-recognition related problems. Along with that, the main-subject of the scene is usually centered at the image plane. Thus, combined with the photographer choice of composing the scene, the tendency of humans of which where to look first a strong implicit variable forming the order of the keyword list. The proposed method exploits this important source of information to predict the objects' position and scale in the image. The paper can be summarized in 3 steps. The first step is to analyze the keywords, in particular word presence in the list, the order and mutual tag proximity, to extract spatial constraints of the objects. The second step is then to model the localization distributions given the extracted features. Each category, given the features, the position of the object and the scale is modeled with a Gaussian mixture. The computation of the mixture parameters is computed by a Mixture Density Network, which is a neural network trained over tag-features and the target parameters (the scaling and the position). The last step is to either incorporate visual cues with the information obtained from keyword-lists about the categories, or rank object detector algorithms. As for the object detection, two window based detectors, DPM and HOG, are used.

3 Conclusions

The innovation that the paper brings, using loosely compiled keywords in object localization, is significant. The proposed method performs significantly better than sliding window based detectors, if a portion of candidate windows are used in both LabelMe and PASCAL VOC datasets. On the other hand, one weakness should be mentioned that the proposed method seems to suffer if the dataset contains similar scenes, meaning that it seems difficult for the method to infer the position and the scales of objects if the dataset shows the same co-occurence distribution. It may result from the fact that it leverages the different object co-occurence distribution in the keyword-list to infer the scales and the positions, hence the scene. Overall, a generic way of modelling visual perception is achieved without even analyzing the perception process; rather human input was used to model this phenomena. It was reassuring after last discussion regarding visual attention and saliency.