Project proposal: From Object detection to Image Captioning in Retail Environment

The goal of object detection and description at a supermarket utilizing computer vision and natural language processing techniques is the issue that we will be looking into. Our proposed pipeline for object detection and description in a supermarket is a challenging and interesting problem that can benefit from a combination of classical image processing and deep learning-based techniques in order to improve the efficiency and accuracy of inventory tracking and management in retail environments.

We will review the literature on object identification, object recognition, spatial reasoning [5], natural language processing and picture captioning to have context and background. We will examine research on convolutional neural networks (CNNs) for object detection and recognition, graph-based methods for spatial reasoning, rule-based and trainable language models for natural language generation and encoder-decoder models for image captioning.

For our pipeline, we will use publicly available datasets such as the Freiburg groceries dataset [1], SKU110K [2] and RP2K [3] dataset for object detection and recognition.

The pipeline for our suggested approach will combine traditional image processing operators, geometric-based algorithms, retrieval algorithms, and deep learning-based elements.

- The images will first be pre-processed using traditional image processing techniques like Canny to detect the edges of the objects we want to identify in the scene.
- Then we will find the patches on the image that contain the objects to feed them to a CNN to classify the single objects. We will also use a retrieval component to enhance the performance of our network.
- After that, we will identify the parts of the image that belong to the classified object, to then perform some graph based global reasoning to find relationships between the objects.
- Finally, we will create natural language descriptions of the scene using rule-based template filling.

To evaluate our results, we will use metrics such as IoU, precision, recall, and F1 score for object detection and recognition, and BLEU or ROUGE scores for natural language generation and image captioning. We will compare our results against existing methods from the literature [4], but we will also try to vary our initial pipeline to find a good mix of classical Computer Vision methods and deep learning algorithms to perform object detection and recognition.

[1]: https://github.com/PhilJd/freiburg_groceries_dataset

[2]: https://github.com/eg4000/SKU110K_CVPR19

[3]: https://www.pinlandata.com/rp2k_dataset/,

[4]: https://arxiv.org/pdf/1904.00853.pdf, https://arxiv.org/pdf/2006.12634.pdf, http://ais.informatik.uni-freiburg.de/publications/papers/jund16groceries.pdf

[5]: https://openaccess.thecvf.com/content_CVPR_2019/papers/Chen_Graph-Based_Global_Reasoning_Networks_CVPR_2019_paper.pdf