Project Milestone

**From Object detection to Image Captioning in Retail Environment**

***Authors***: Bucciarelli Davide <271102@studenti.unimore.it>,

Fino Alessandro <313424@studenti.unimore.it>,

Pavone Cosimo <272636@studenti.unimore.it>

***Introduction:***We will study the problem of object detection and description at a supermarket using computer vision and natural language processing methods. Our plan for supermarket object detection and description consists of detecting the objects in the scene and provide a description of them. This approach can benefit from a combination of traditional image processing and deep learning-based techniques to increase the effectiveness and precision of inventory tracking and management in retail settings.

***Problem statement:***  
For our pipeline, we will use publicly available datasets such as the Freiburg groceries dataset [1], SKU110K [2], RP2K [3] and Grocery Products dataset [4] for object detection and recognition. In the first phase of the project, we started working on the last-mentioned dataset using traditional image processing techniques like Canny to detect the edges of the objects we want to identify in the scene. The dataset contains 8350 training images of grocery products, organized in 80 hierarchical classes, and 680 annotated test images of supermarket shelves. In general, our overall plan can be described as:

* 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 perform single object detection over self-acquired images to then use a retrieval component to find grocery products that are similar to the subject of the image.
* 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.

***Technical approach:***   
Regarding multiple object detection for finding items on retail shelves, we decided to use an approach based on deep learning since we found out that classical methods are not feasible for this use case, due to the complexity of the scene, in which we have many objects and borders that we don’t want to detect.

#immagine della detection schifosa

To do this we want to use a pretrained convolutional neural network and fine tune it for our specific task, such as YOLO or FasterRCNN.

On the other hand, for single object detection, classical methods are good enough. In particular we tried several different approaches and the one which gave the best results was the following:

* We applied a gaussian filter to smooth out the image and remove some of the noise.
* We then split the R, G, B channels of the images in order to treat them separately.
* Then, a multiresolution pyramid has been built for each channel to detect all the edges strong enough to be found in at least one layer of the pyramid.
* The Canny edge detection algorithm is applied to each layer of the three pyramids to detect borders in each of the R, G, B channels of the image to find most of the edges.
* The edge maps are then merged and binarized using thresholding.
* We used dilation and erosion to try to connect into contours the edges that do not form a perfectly closed line.
* Contours are then detected from the closed edges.
* We then find the biggest contour and draw a bounding box around it.

#Immagini singole dei result

***Intermediate/Preliminary result:***

To evaluate the results of our single object detection approach we used the IoU [] metric.

##mettere la formula in latex

We acquired and annotated a small dataset of 87 images of groceries in various light conditions and with different, plain colour, backgrounds for measuring our performance [vgg annotator]. We used an automatic testing script to find the best parameters for our task. In particular, we focused on finding the best values for THL and THH for the Canny Edge detection algorithm. We passed through our sample set several times changing those values, in particular:

* We used a fixed ratio between THH and THL, trying values from THL=50 to the maximum valid value for the given ratio.
* Then we repeated this process for ratios starting from THH = 1.25\*THL to THH = 2.5\*THL using steps of 0.25.

These are our current results:

##immagini di tutti i plots//

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

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

[3]: <https://www.pinlandata.com/rp2k_dataset/>

[4]: Marian George, Christian Floerkemeier, "Recognizing Products: A Per-Exemplar Multi-Label Image Classification Approach", ECCV 2014 <http://www.vs.inf.ethz.ch/publ/papers/mageorge_products_eccv2014.pdf>

[5]: <https://arxiv.org/pdf/1810.01733.pdf>

[6]: <https://arxiv.org/pdf/1904.00853.pdf>

[7]: <https://arxiv.org/pdf/2006.12634.pdf>

[8]: <http://ais.informatik.uni-freiburg.de/publications/papers/jund16groceries.pdf>

[9]: <https://openaccess.thecvf.com/content_CVPR_2019/papers/Chen_Graph-Based_Global_Reasoning_Networks_CVPR_2019_paper.pdf>  
[10]: J. Canny, “A computational approach to edge detection,” IEEE Trans.Pattern Anal. Mach. Intell., vol. 8, no. 6, p. 679–698, 1986  
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.420.3300&rep=rep1&type=pdf