2nd problem. We are given a dataset that contains an unlabeled set of images and a smaller dataset that contains labelled images. Those images represent shirts from an e-commerce, so usually the background is plain and the product is the focus of the image. The unlabeled dataset provided contains 5K samples and the labelled one contains just 10 images, all of **blue** shirts.

We would like you to analyze the datasets and build a model that is able to retrieve all the blue shirts from the 5K dataset. For this purpose, the use of open source libraries is encouraged.

You can download the datasets from http://app.goldenspear.com/shirts.tar.gz

1. **Problem analysis.** Write down your solution proposal.

   My approach consists in the implementation of a program (color-detector.py) that helps us identify the range of HSV values corresponding to the blue color (based on the labeled dataset), and a model (main.py) that detects those regions in the input images that fall within that range of color.

   Firstly, all images are resized to the same width and height so that it is possible to define a threshold that is valid for every image. This threshold is used by the model to determine whether or not an image has a relatively large region of pixels which color falls within the range previously computed. If the images are not resized the threshold would have different meanings depending on the size. In other words, what could be a very large region for some images might not that big for others.

   Furthermore, as you can see in Figure 3, the model considers the sum of all blue contours detected, that are big enough (contour area > 1000), to determine whether the input image satisfies the threshold requirement of 9000.

Figure 3: *Examples of blue shirts detected based on multiple contours.*

Finally, I am aware that this approach entails some issues. For example, the model has no way of knowing whether the blue region detected corresponds to a shirt or something else. Additionally, resizing the input images implies deformations in the image that might cause a loss of proportions. Also, the model does not detect those blue regions which pixels are not within the range found based on the labeled shirts.

2. **Models validation.** As we don't have validation samples on the dataset, find a way to visually demonstrate the capacity of the model.

   For this problem and with my current implementation, I cannot think of a way to validate the model that is not one by one checking whether or not the recognized images are the correct ones. Unless, we had another dataset with labeled images and we could test the model with it.

3. **Final summary**. Write down what would you have done if we had given you more time and data.

   An alternative to finding the HSV values manually might be the application of a Genetic Algorithm (GA) designed to learn the HSV values corresponding to different types of blue. Although, in this case, the definition of the fitness function is a bit challenging as it should maximize the white region in the mask generated by the program, but being careful to not end up with the range of HSV values that includes all the colors.