COMPUTER VISION COURSE

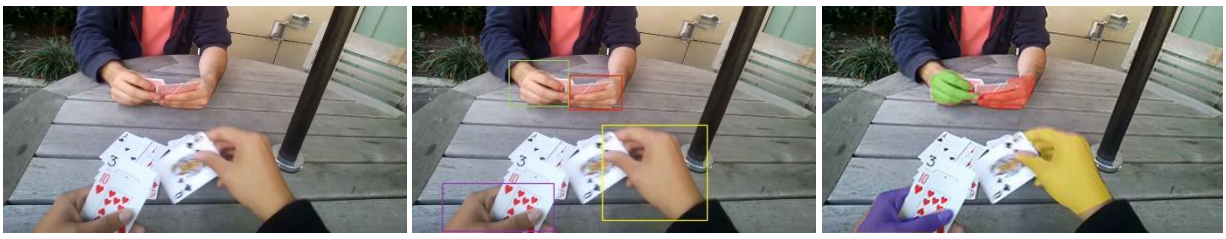
Summer project 2022

Human hands detection and segmentation

Introduction

Human hands play a very important role when people interact with each other or with objects and tools. Therefore, the reliable detection of human hands from images is a very attractive result for applications of human-robot interaction, gesture recognition or human activity analysis.

The goal of this project is to develop a system capable of 1) detecting human hands in an input image and 2) segmenting human hands in the image from the background (i.e., binary segmentation) – see examples in the images below. The appearance of hands can change sensibly from image to image, depending on the viewpoint, and can be influenced by many factors such as skin tones or the presence of clothes (e.g., shirt, rings) and objects.



Example of the required outputs. From left to right: input image, human hands detection and hands segmentation.

The output format shall be:

- hand detection module: the original image with a bounding box for each hand recognized in the image;
- segmentation module: the original image where pixels belonging to a hand are given a dedicated color (a different color for each hand) – you may use transparency (as in the example above) or substitute the original color completely.

While developing the project, please consider that:

- Images without human hands shall be correctly analyzed
- Multiple hands can appear in an image
- Hands can appear from any viewpoint, with different scales and skin tones

Test dataset

To assess the performance and robustness of your system a benchmark dataset with annotations is provided at the following link:

https://drive.google.com/drive/folders/1ORmMRRxfLHGLKgqHG-1PKx1ZUCIAJYoa?usp=sharing

The README file provides the description of the dataset.

The dataset contains images categorized by level of difficulty (details in the README):

a) Images with similar backgrounds where few hands are present and clearly visible;
b) Images with different backgrounds and many hands present, possibly with partial occlusions;
c) General hand pictures of people of different skin tone and gender.

The number of images on which your system will be evaluated is as follows:

- If you develop your project alone, your system will be evaluated on the subset a) of the dataset;
- If you work in a group of two students, your system will be evaluated on the subsets a) and b);
- If you work in a group of three students, your system will be evaluated on the entire dataset.

Performance measurement

You should evaluate the performance of your approach considering:

- For hand detection, the Intersection over Union[1] (IoU) metric;
- For hand segmentation, pixel accuracy[2] for both classes (hand, non-hand).

The metrics mentioned above need to compare the output of your system against the ground truth, namely what is considered "the truth" for each output image. Ground truth annotations are already included in the dataset provided for the final evaluation of your system.

The ground truth provided is needed to assess the performance of your system. If your solution relies on machine learning/deep learning (which is not needed – just your choice), you will need the ground truth of a sensibly larger dataset. Collecting additional ground truth is a time-consuming task, but you are free to share the output of the ground truth with other colleagues belonging to other groups. In other words, you are allowed to cooperate with everyone to share the ground truth collection. The organization of the ground truth collection is completely free; if you wish, you can use the project section in the moodle forum.

If you collect your own ground truth, you should exploit the same standard already used in the benchmark dataset and detailed in the README file. If you use additional dataset(s) and related ground truth, **you must describe this in the report**, **providing a link** to an online resource (e.g. UniPD google drive) where such additional dataset is available.

If you use machine learning/deep learning in your project, please consider that **you must not use any image coming from the test dataset** in your training/validation dataset.

[1] https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/
[2] https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2

Project delivery

The project must be developed in C++ based on the OpenCV library. The only allowed exception is the usage of Python code for developing the networks if you decide to exploit this family of techniques. However, **the project must compile (using CMake) on the T.2020 virtual machine with the current software and library installation**. The trained deep network shall be provided as a file.

You need to deliver your project including:

- All the source code (both C++ and Python);
- CMake configuration files (the use of CMake is mandatory);
- A report (no page limit) presenting your approach and the performance measurement on a set of images detailed below.

Working in group, you should **clearly identify** the contribution of each member in terms of ideas, implementation, tests and performance measurement. You can organize the work as you prefer: you are not forced to assign one specific step to each group member. Please also include **the number of working hours** per person in the report. This is needed for a monitoring on our side on the effort requested – the evaluation will not depend at all on the number of working hours, but on the quality of the result.

Your report **must include** the results of the hand detection and hand segmentation (both output images and metrics values) for the test images in the provided benchmark dataset (according to the number of members of your group).

As a reference, the images in the provided benchmark dataset have been sampled from the following publicly available datasets:

- EgoHands: http://vision.soic.indiana.edu/projects/egohands/
- HandOverFace(HOF): https://drive.google.com/file/d/1hHUvINGICvOGcaDgA5zMbzAIUv7ewDd3

Feel free to add more test images, taken from the internet or acquired using your own camera.