The researcher's challenge

The aim of this project is to build a human-computer interaction system using hand gestures control based on inputs from an RGB camera. In this report, we analyse several problems an algorithm that aims to solve this might face and possible difficulties with implementations and we propose a sketch of the development plan.

Controlling a device simply by using hand gestures is something that humanity has been thinking of for a long time, especially in Sci-Fi movies. Given today's advances in technology, however, this endeavour does not seem that far-fetched. In the next 6 months, our team is going to design one such system that works on desktop computers as an SDK.

One possible approach for this consists of multiple steps: first, the hand needs to be detected in the picture, then motion detection should be performed on the last frames in order to detect the gesture and lastly, the computer should take a decision and execute an action after receiving the gesture.

The first challenge is represented by the diversity of the possible inputs. The colors of the image, the proximity of the hand to the camera, the speed of the move, the orientation of the person in front of the camera, the background of the image, which might consist of walls with pictures, furniture, parts of human faces or bodies, all make this problem incredibly hard. Because we have chosen to develop a solution only for desktop computers for now, the orientation is somehow limited, as we expect the person to be in front of the device, whereas a mobile phone would have given the user more freedom.

In order to minimize the impact that the light or skin color can have on the input images, the color space that should be used is YCbCr, which separates the brightness component from chroma. To detect the hand, the background should be blurred and then removed, the colors should be filtered (possibly binarized), so that the solution is invariant to skin color. The next step is tracking the movement of the hand in order to understand the gesture shown. One possible approach for this could be using a convolutional neural network for embedding the pre-processed frames into a fixed-size array of real

numbers and then feeding several consecutive frames (possibly of variable length) to a sequence model such as a vanilla recurrent neural network or long-short term memory that is trained to predict the correct gesture. Finally, the decision-taking component (what should the system do after the gesture has been recognized ?) can be modelled as a Markov decision process in which the reward is the negative time for the user to perform the desired action.

One of the implementation challenges is that this entire process of detecting where the hand is situated (if there is one in the image) and then tracking its movement in order to classify the gesture is very computationally intensive, while one of the first requirements of on such system is to run in real-time, so a compromise should be made between performance and speed. The pipeline might be trained offline in order to make it faster. Also, because we are using a supervised learning approach, a dataset should be provided and its collection requires manual labor. Another problem is the variety of hardware devices that the application should be compatible with (has to be compatible with multiple operating systems, the camera specifications might lead to different image quality, different FPS etc.)

Because there is not much time available, the pipeline for pre-processing the images and hand motion detection should be implemented in the first month, while at the same time manual workers will be paid to collect the necessary data for training the model. Another month can be dedicated for the decision component. After having the first working prototype of this pipeline, the main effort can be put into adjusting the pieces to have better performance and better speed and intensively testing the product before making it available to the public.