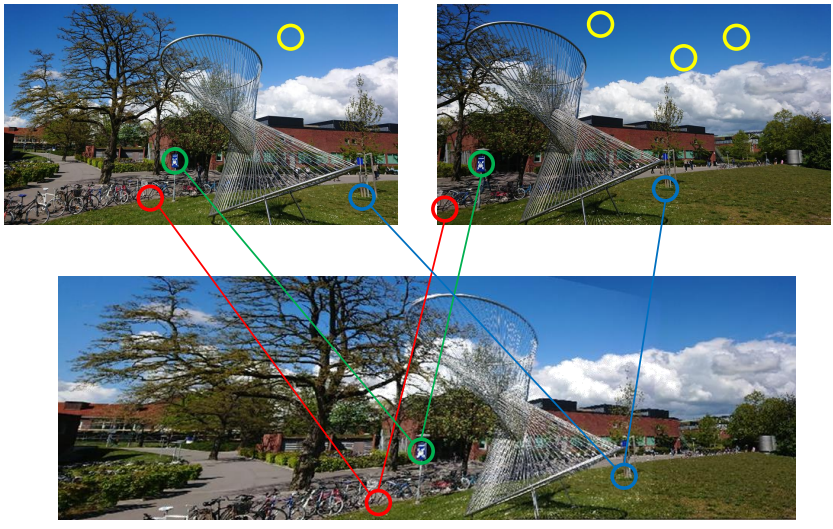
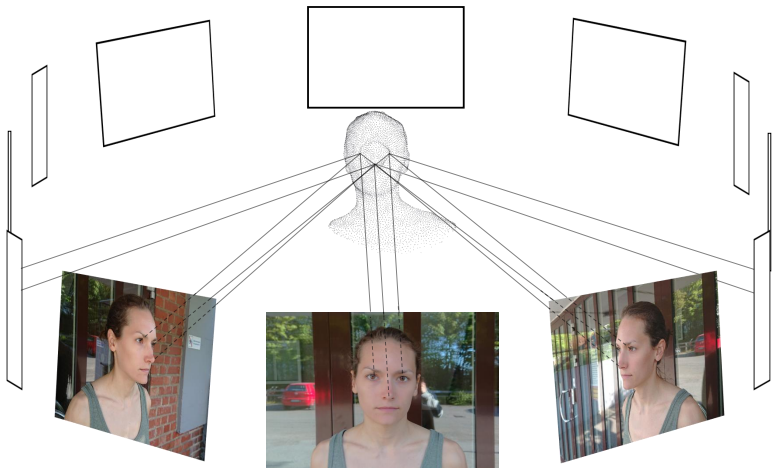
Popular Summary

Can computers understand the world as well as humans do? Can a robot navigate on its own? Can a drone remember the appearance of a room? And can you from a number of images learn what the environment they depict looks like? All of these questions have a connection to the contents of this thesis.



The figure shows how to combine two different images of the same scene into a panoramic image. This requires the identification of interesting and matching points in both images, for example marked by the red, green and blue circles. The yellow circles show an example of points that are bad to use, since these are not unique.

Computer vision is an area which focuses on teaching computers how to gain knowledge and understanding from images, just as humans do when they see something. In many ways cameras are similar to the human eye, and humans and other animals are in general very good at understanding which type of objects that are in front of them or how far away different things are. The idea with computer vision is that computers should be able to do this as well. One thing that images can be used for is to create *3D models* or *maps* of the world. We humans can determine the depth of what we see and using this, our memory and our experiences we can create our own map of, for example, a room or a flat. The same thing can actually be achieved digitally, using computers. The procedure is similar to that of creating panoramic images. Today, most mobile phones have both a camera and an application for creating panoramic images – that is, many small images stitched together to one larger image. The essential parts of image stitching is to find *interesting* points that can be seen in both images and then to move the images such that these points match. An example of this can be seen in the figure above.

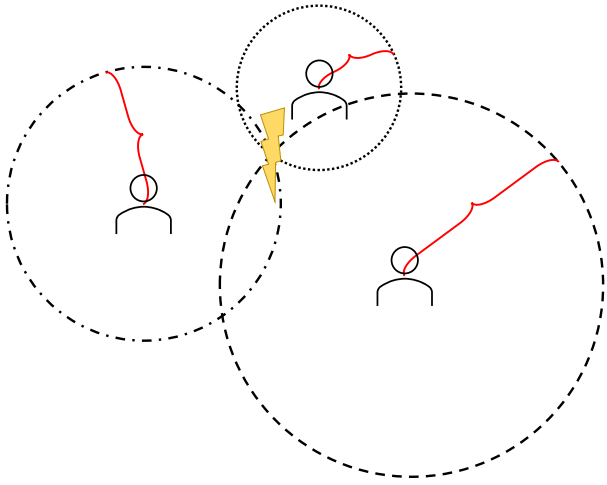


Using several images the 3D position of the "interesting" points can be found. If there are enough images one can create a 3D model, here represented as a point cloud, of the person in the images.

Similarly, with enough images taken from different angles, these can be "stitched" into a 3D model. The depth can be recovered since the images are taken at different positions, as can be seen in the figure above. To estimate 3D coordinates using interesting image points and known camera positions is called *triangulation* and if the camera positions are unknown but estimated as well, we call it *structure from motion*. If sufficiently many points are triangulated a map of the environment is obtained.

Such maps can also be created using other types of sensors and signals, such as microphones and sound. Sound describes some sort of information in one dimension in the same way that images do in two dimensions. Most people have probably, at some point, counted the seconds from the flash of a lightning until the thunderclap can be heard. Through this, you know how far away the thunderstorm is. You do not know in which direction it is located, but you do know that it is positioned somewhere on a circle where you are the centre, and the counted distance is the radius. If you could also know the thunder's distance to two other positions, you could draw two more circles and then the thunderstorm would be located where these intersect. This has been illustrated in the figure on the next page. The same procedure can be used for microphones and loudspeakers that are set up in a room, and if there are sufficiently many it is actually enough to only know the distances between them to compute the relative positions of both the microphones and the speakers. This gives a map of the microphone positions, in many ways similar to the map that can be created from images.

Once the 3D maps are obtained, they can be used for *positioning* – to find out where in the



The image illustrates how you can find out where a thunderstorm is, if three people hear the thunder from different positions. By drawing a circle around each person, the position of the storm is given by the intersection of the circles.

map you are, either using sound or images. One example of a system that uses signals that are similar to sound is GPS, which is used a lot for outdoor positioning and navigation. The GPS does, however, work less well indoors, and because of this it can be good to have other systems that can be used in similar ways for indoor positioning. For the positioning to be as good as possible it is important that the map is as good as possible and in general the maps get better if more measurements are used (for example more images). Therefore, it is a good thing to be able to merge different maps of the same environment, to a more exact map and to be able to update it in case something in the environment changes. This can be done by identifying corresponding interesting points in the different maps and then stitching them so that they coincide – just as for panoramic stitching and triangulation.

Map merging can, for example, be useful for self-driving cars. Today, many cars have one or several cameras which can be used to determine the position of the car in an already known environment. Imagine that we have a map of a city and that a car is driving through this city. While driving, it will collect many images and using these images it can create its own, *local* map of the city. Now, if the local map can be added to the large, *global* map, the updated global map will after that contain more information than before. This makes it more exact. So if all cars that drive through the city can do the same thing, the map will gradually get better. Also, if some infrastructure of the city is changed, this will also be captured in the map, without the map being re-created.

In this thesis we first focus on finding good measurements between microphones and loudspeakers. We then cover the topic of creating maps using such measurements and the more exact the measurements are, the better will the resulting maps be. Finally, we show how

several such maps – consisting of 3D point clouds – can be fused into a single, more accurate map. The local maps can be created either using sound or images, as in the examples above.