EECS C106B / 206B Robotic Manipulation and Interaction **Spring 2020**

Lecture 10: Chapters 1 and 2 of MaSKS

Scribes: Ritika Srivastava and Riddhi Bagadiaa

10.1 Perception and Vision

Vision and perception is an extremely important part of Robotics. As human beings, processing images and videos takes up a large part of our brain capacity. However, visual capabilities of a robot are still very limited and is an unsolved problem.

The Fundamental Problem
The basic Vision problem is to take multiple 2D images from different perspectives and mark out the same "features" in each image. Then use mathematical principles on these images as input to find the camera pose, calibration and scene structure representation. This output has several applications.

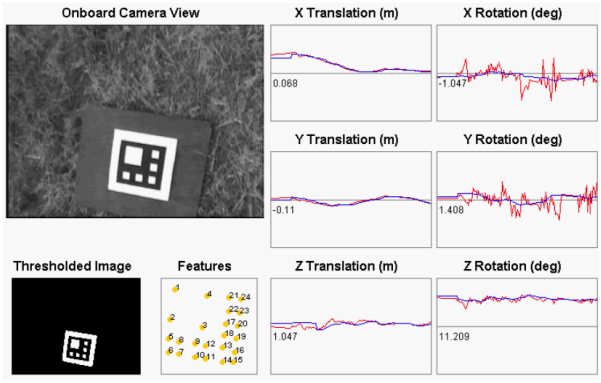
10.2 Applications

10.2.1 Autonomous Highway Vehicles

The Berkeley PATH project started in the 90s is still active, and has evolved immensely since.

10.2.2 Unmanned Aerial Vehicles

One application was trying to land a helicopter on a moving battleship. It requires a sensor and a closed loop feedback system which continuously tracks the target and adjusts itself. Other applications include having a fleet of drones to perform drone shows and using drones in agriculture since they can move through dense paths.

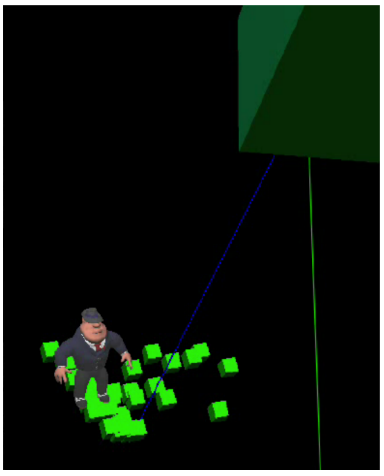


10.2.3 Real Time Virtual Object Insertion

This is very similar to Augmented Reality. An object is placed on an image. The object is calibrated such that it remains in the same position relative to the viewer, even when the image is moved. The procedure is as follows:

Feature points are marked on the image and tracked. This gives the depth and pose of the camera.

The object is placed on the image with respect to those feature points.



The image is removed from underneath but the object remains in position, anchored to those feature points.

An application of this is Virtual Advertising. Many online streaming sites process the video and simultaneously impose virtual objects onto the video. For example in Football games, multiple pictures of the stadium

are taken from various angles, to understand the geometry of the stadium. These images are processed and different ground markings are superimposed on the images, for viewers' convenience.

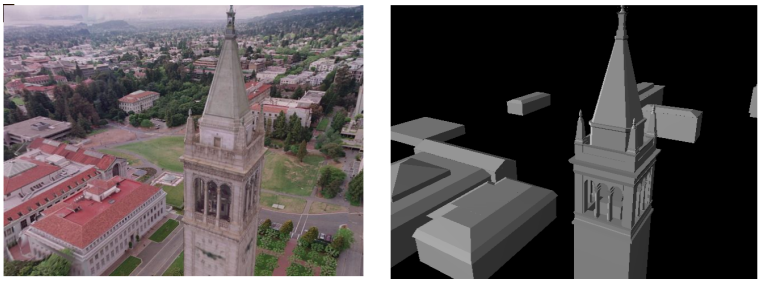
In this image, the yellow line has been placed virtually for the benefit of the viewers.

10.2.4 Virtual Reality

Similar to the previous application, images can be captured and edited such that their backgrounds are removed. These images can then be placed in an alternate background to give similar effects as Virtual Reality.

10.2.5 Image Based Modelling and Rendering

Multiple images are taken of the same object. Corresponding feature points are matched in each picture. This helps model the object in a software. Using that model, different images of the object can be generated.



This image shows the model being created using the image of the Campanile.

An (obsolete) application of this technology is to create virtual backgrounds for movie sets.

10.2.6 Image Alignment, Mosaicing and Morphing

This technique is used to create one large picture using multiple smaller pictures. Once again, pictures are taken from different angles and the common points in each image are matched. These matching features are used to stitch the images together to make a continuous looking picture.

An application of this is in Google Maps street view, where the streets are seen continuously as if it's one large picture.

10.3 General Steps

10.3.1 Feature Selection and Correspondence

This is a general summary on how to perform feature selection and correspondence, as mentioned in the applications above.

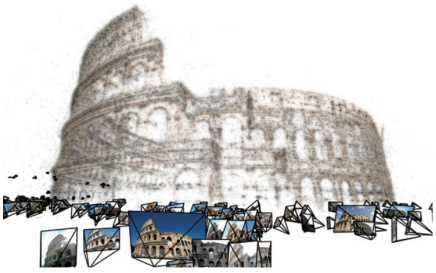
1. Take pictures of the object from different angles and views

2. Mark feature points that are easily distinguishable by sensors. E.g. in a picture of a building, building corners, window corners or other objects placed in front of the building would serve as good feature points. There is a trade off between having a small baseline and large baseline while feature picking. Baseline refers to the distance between the two cameras. Wider the baseline, the easier it is to triangulate between the point and the camera. Smaller the baseline, easier is the correspondence.

3. Structure motion and recovery - plot the points based on the features. Once the position of the camera is known, triangulation can be done. This is used to render and recover the 3D geometry of the object, and perform depth estimation. Typically, points are estimated from regions with less points to a=regions with more densely populated points. The object can be recreated by borrowing pixel values from original image.

10.3.2 Image-based 3D Modeling

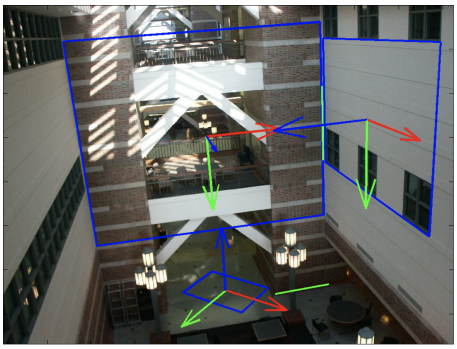
Once you have more faster and better computer processing power and speeds, you can use SIFT feature detector and build 3D models. Steve Seitz and lab took images of the Rome and followed the same pipeline that existed 20 years ago. With more computation power it was possible do the tracking and correspondence, they were able to create the 3D model of Roman sites. This is done with a point cloud.



This is the same technology behind Google Street View and other Apple and Microsoft projects.

10.3.3 Image-based 3D Modeling

There is a smaller end of the spectrum. Without the high computation power used to recreate Rome, it is possible to handle the same problem on a smaller scale, such as on a cell phone.



To pursue this problem it is important to understand that the world is not a random point. There are 3D structures that can be perceived from images. We want out robots to be able to find this information from a stereo system. Once you are able to model the geometry of the image is is possible to add to that geometry.

There is a color difference, but the geometry is the same. This is a more impressive version of augmented reality than the man on a newspaper.

With a single image it is possible for you to understand the geometry of a object in am image and the location of view relative to that object.

With the understanding that the world is not a point cloud, can we convert the point clouds into something regular? It is possible! The Holicity dataset shows $20km^2$ of downtown London. This includes a bird's eye view of a CAD model of the city, the viewpoint coverage, panoramas, RGB images, and rendering with surface segments and depths.

This is a continuing problem.

10.3.4 From Image to CAD models

Now it is possible for us to build CAD-like models using those images to create high level geometry CADS which incorporate principles of end-to-end learning and geometric structure.

10.3.5 Evolution of Interface and Media

With civilization, people started with 1D media. The Inca used knots called Quipu to keep track of information. Then people continued to 2D media with stone carvings, paper, or touch screen notes. We are still stuck with 2D media. These days technology has been moving to 3D media with virtual reality (Vive, HTC, Hololens). In the future all of your information will be integrated into 3D mediums.

10.3.6 Examples

You can get pretty high quality renderings.

It is possible to create an immersive experience. In an environment where the technology fails to adapt, it is possible to alter the environment so that your technology works.

View immersive experience by clicking here

Rome wasn't build in a day! But a digital Rome may be built in a day!