BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Visual Camera Orientation Estimation using Machine Learning

Martin Kubička*

Abstract

The purpose of this work is to create a model using spherical convolutional neural networks that can estimate the orientation of a camera from two inputs, where the first input is a panorama and the second input is a photograph capturing a specific part of the panorama. In other words, the task is to find where in the panorama, which is the first input, is located the photo, which is the second input. In addition to three created models that address this problem, six new datasets have also been created, which expand the currently available number of datasets whose photos are in equirectangular or stereographic format.

*xkubic45@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

We often encounter photographs and videos not only in physical form but also on the internet, where it may seem that there is an inexhaustible number of these images. Most of us, when looking at a photo, may wonder about various questions such as where the photo was taken, but we usually don't ask ourselves about the orientation of the photo. We don't ask this question because, as humans, it's more or less clear to us at first glance, but to computer vision, it's not so obvious. Knowledge of camera orientation is a fundamental element of computer vision. Precise estimation of the camera orientation is crucial for many applications. A common example is augmented reality, where we can insert various objects into real space, thus truly expanding reality. Augmented reality has many educational, military training, and entertainment applications [1].

2. Task

Task is to estimate camera orientation using machine learning methods, specifically using spherical convolutional neural networks. This task involves creating a dataset and acquiring other datasets for and then developing architecture for training. The orientation of the camera in this work refers where in 360 degree panorama is located another photo, so there are 2 inputs. Camera orientation is defined by three angles: pitch which is rotation around X axis, yaw which is rotation around Y axis and roll which is rotation around Z axis. So our task is from the two mentioned inputs, estimate the three mentioned angles. Visualized task pipeline can be is shown in Figure 1.

3. Spherical CNNs

With this type of convolution, we talk mainly about two of their advantages. The first advantage is equivariance, or invariance to rotations, and the second advantage is the significant deformation that occurs when we want to project a sphere onto a 2D surface. These properties are typically not found in ordinary convolutions.

The principle of rotation equivariance or invariance is similar to ordinary convolutions, which are equivariant or invariant to translations. In practice, this means that if a convolution is not equivariant or invariant to rotations, then when we display the output of some photo after convolution and then display the output of the same photo but rotated around the Z-axis, the outputs will not resemble each other, and thus will look completely different. Therefore, if a spherical convolution ensures rotation equivariance, the outputs will be the same, where the second output will be rotated around the Z-axis by a certain number of degrees, so if we rotate the second output back by the same number of degrees, both outputs will match. We want equivariance to rotations when we want to preserve information about this angle. Rotation invariance means that both outputs will look the same, and in practice, it works in such a way that the filters rotate, and that is the reason why the outputs will look the same. As we might guess, invariance loses information about this angle. Rotation equivariance is shown in Figure 2 and rotation invariance is shown in Figure 3.

Another mentioned advantage is that if we want to project a sphere onto a 2D surface, deformations will occur, which can be a problem when training using ordinary convolutions. In reality, the input for these convolutions is a 2D photo onto which a sphere is projected using sphere to 2D plane projections like equirectangular projection and spherical convolutions used special size of filers to handle this. Distortions caused by projection can be seen in Figure 4 and visualization how spherical convolutions pass through photo can be seen in Figure 5 which was taken from [2].

4. Results

For PACNoRoll/GeoPose3K dataset it was possible to estimate the pitch angle with an accuracy of 39/57% if we tolerate an error of 10 degrees, 53/80% if we tolerate an error of 15 degrees and 61/91% if we tolerate an error of 20 degrees and for yaw angle, if we tolerate and an error of 10 degrees accuracy is 6/16%, if we tolerate an error of 15 degrees accuracy is 10/20% and if we tolerate and error of 20 degrees accuracy is 16/24%. Visualization of prediction compared to ground truth can be seen in Figure 6.

5. Conclusion

Several models were created for resolving this task, where each has a different idea behind it and brings various results. The best results were shown by the direction of semantic segmentation, which brought the best results, where for each angle, a separate model is created where initially, the pitch angle is estimated and then the yaw angle in case we are estimating only these 2 angles. During the development of the model architecture, we encountered the limits of current hardware, specifically the size of GPU memory. Another challenge was the fact that there are no pre-trained models for this type of convolution, so we trained from scratch.

In addition to the created models, several datasets were developed, which are a significant contribution, as there are not many datasets that contain panoramas in equirectangular format and even fewer datasets that contain panoramas in stereographic format. Part of the work also includes programs that generate these datasets, and thus they can be expanded.

Acknowledgements

I would like to thank my supervisor doc. Ing. Martin Čadík Ph.D. and my family for guiding me through the process of creating this thesis, giving me advice along the way and passing on their knowledge.

References

[1] Meng Xu, Youchen Wang, Bin Xu, Jun Zhang, Jian Ren, Stefan Poslad, and Pengfei Xu. A critical analysis of image-based camera pose estimation techniques, 2022.

[2] Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical cnns, 2018.