Learning the Depths of Moving People by Watching Frozen People

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

October 8, 2019

While humans are capable of maintaining a logical interpretation of depth in an environment even when both objects observed and the observer are moving, it has been difficult to recover depth-information even from a stationary, handheld camera observing a dynamic scene. Li et al. [1] propose a method to predict dense-depth in environments where the camera is naturally moving (as it is handheld) and observing a dynamic environment of people moving freely. In order to learn priors for human depth from data, they present a new dataset which they refer to as the *MannequinChallenge* (MC) dataset. The data was collected from videos uploaded to YouTube as part of the so-called *Mannequin Challenge* where all people in a location would freeze and try to stand as still as possible. This means the entire scene is stationary so camera poses and depth can be estimated and used as training data.

Based on this training data Li et al. train a neural network which, given an image, a binary mask of human regions, a depth map of the environment (non-human regions), a confidence map and an optional human keypoint map, outputs a dense depth map over the entire image. The depth map of the environment is computed using two frames of the video by first estimating the optical flow field using FlowNet 2.0 and then using the relative camera poses and Plane-Plus-Parallax representation. As optical flow can often be noisy, especially in internet video clips with often challenging lighting and recording conditions, the authors provide a confidence map which estimates which regions of the depth map are of high-confidence and which regions are low-confidence.

The authors evaluate their model quantitatively and qualitatively on the MC dataset, on a subset of the TUM RGBD as well as internet videos of dynamic scenes. They find that their model performs well even in dynamic scenes despite having been trained on people "frozen" in scene, outperforming previous models. However, as Li et al. point out the approach still has its limitations, as it requires known camera poses, which may not always be available, and may not be accurate for scenes with non-human moving objects (i.e. cars).

References

[1] Li, Zhengqi, et al. "Learning the Depths of Moving People by Watching Frozen People." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.