Learning Optical Flow from Real Robot Data

Parth Shah (pshah9@stanford.edu)

Motivation

Manipulation and grasping is a very challenging task in the field of robotics. Current state of the art techniques requires mesh models of the object. I hope to extend recent advances in robot tracking [1] to enable robots to learn optical flow. Optical flow is one way a robot can learn the dynamics of a scene without mesh models.

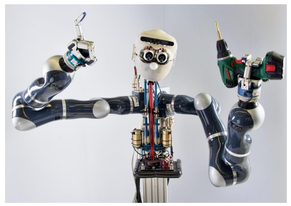


Fig. 1 Apollo – robot platform utilized

Data

- All available ground truth datasets for optical flow are either not densely annotated (i.e. KITTI [4]) or are synthetically produce (MPI-Sintel[5], FlyingChairs[3])
- I will generate my own dataset, from real robot data, to include phenomena that can not be modeled synthetically
- **Input -** rosbags with RGB-D images and joint encoder values
- **Steps**
  1. Track
  2. Render
  3. Mask
  4. Calculate flows
- **Output**
  - 4300 masked RGB images pairs with corresponding flow values

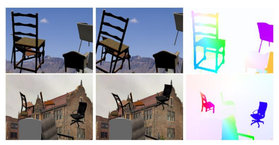


Fig. 2 FlyingChairs dataset

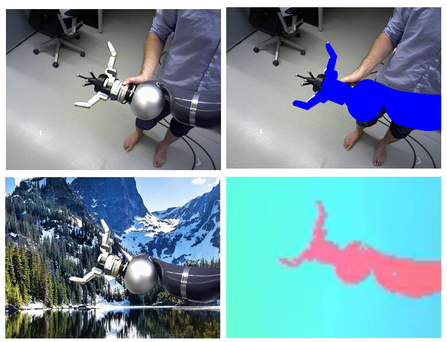


Fig. 3 Raw input image (TL), RGB image with rendering (TR), masked RGB image (BL), output optical flow (BR)

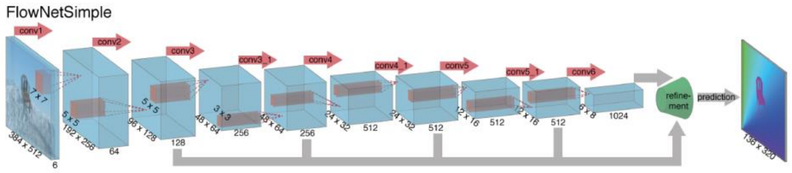
Models

- **Baseline – Lucas & Kanade [2]**
  A differential technique for estimating optical flow. It relies on partial derivatives and assumes that displacement is small and approximately constant.

$$I_x(q_n)V_x + I_y(q_n)V_y = -I_t(q_n)$$

Eq. 1 The central equation that Lucas-Kanade is built off

- **Convolutional Neural Network – FlowNet [3]**



FlowNetSimple consists of two halfs – one focused on extraction, followed by emphasis on refinement

- Extraction – A series of 9 convolutional layers with intermediate leaky ReLU activation layers. Works the input images down from [384, 512, 6] to [6, 8, 1024].
- Refinement – 6 iterations of upconvolutions, convolutions + image resizing, and concatenations to produce optical flow prediction at the original resolution

Notes: Upconvolution is the opposite of a convolution. Image resizing utilized bilinear interpolation.

References

[1] Cristina Garcia Cifuentes and Jan Issac and Manuel Wthrich and Stefan Schaal and Jeannette Bohg. Probailistic Articulated Real-Time Tracking for Robot Manipulation. Robotics and Automation Letters. 2017.

[2] Lucas, B., and Kanade, I". 1981. An iterative image registration technique with an application to stereo vision. Pro~ DARPA lmage Understanding Workshop,

[3] Fischer, Philipp, et al. "Flownet: Learning optical flow with convolutional networks." *arXiv preprint arXiv:1504.06852* (2015).

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR), 2013.

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, ECCV,

Features

Input features are 2 sequential RGB images [584, 312, 6]
Derived features are

$$\sum_{n=1}^{N=\#\ of\ convs}(layers_n) * (kernel\_size_n)^2\ )$$

Results

- **Loss** – AEE (average endpoint error) in pixels

Technique	Training (AEE)	Test (AEE)
Lucas-Kanade	-	7.94
FlowNetS (50)	1.87	14.86

Table 1 Displays the results of the two techniques implemented

- **Visualization of Output**

Ground truth
AEE = 0.0

FlowNetS
AEE = 0.39

Image N/A – flow values ranged from $[-10^2, 10^2]$ distorting color wheel

Lucas-Kanade
AEE = 13.53

Discussion

Due to limited availability of computational resources the network was only trained on a subset of the training data (50 of 3450). The results, as visualized above, show promise, but in order to bring down the test error the network needs to be trained on the entire training set to account for different variations present in the data.

Future Work

- Train network with the entire training dataset (3450)
- Introduce challenging phenomena like occlusions
- Benchmark against known datasets (MPI-Sintel [5] and FlyingChairs [3]) for performance comparisons