CS231A
Computer Vision: From 3D Reconstruction to Recognition
Author: Luke Jaffe
Due: 02/01/2018

Project Proposal

I. Background

FS Studios, a software contracting company which focus in-part on computer vision applications, has recently started a project involving human-computer interaction with a ToF camera produced by PMD Technologies. This camera, the CamBoard pico monstar, collects grayscale + depth (G-D) data at up to 60 fps. This project will involve an informal collaboration with FS Studios to produce algorithms and software specialized for data from this camera.

II. Problem Statement

The first part of the project will focus on human hand pose estimation and basic hand action recognition from public RGB-D or G-D video data. Once an effective solution is developed, this solution will be applied to video data from the pico monstar camera. This problem is interesting for a variety of applications involving human-computer interaction including video games, virtual/mixed reality, and robot imitation learning [4].

III. Data

1. Open-Source Datasets

Multiple large datasets which contain annotated 3D hand video data have been identified. Some of the most prominent datasets are listed below. The listed datasets all have hand-joint annotations of RGB-D data. The First-Person Hand Action Dataset also has hand action labels from 45 categories. This dataset is the most likely candidate for this study since we intend to explore hand action recognition in addition to pose estimation.

- NYU Hand Pose Dataset [3]
  - 80k frames of RGB-D data with binary segmentation annotations, and 42-DOF annotations
- BigHand2.2M Dataset [2]
  - 2.2M frames of RGB-D data with 21-joint annotations
- First-Person Hand Action Dataset [4]
  - 105k frames of RGB-D data with 21-joint annotations and 45 different hand action labels

2. Manually-Collected Data

G-D video data will be collected from the pico monstar camera to test algorithms developed from the above open-source datasets. Given time constraints, only a small amount of data will be gathered and annotated. This data will likely be annotated only with bounding boxes around the hands, in addition to hand action category ('writing', 'typing', etc.).

IV. Methods

To the author's knowledge, all modern methods for hand pose estimation and hand action recognition utilize convolutional neural networks (CNNs) [1, 2, 3, 4, 5, 6, 7]. The BigHand2.2M paper [2] and the First-Person Hand Action paper cite *Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation* [5] as the state-of-the-art for pose estimation. This method uses a spatial attention mechanism and Particle Swarm Optimization (PSO) to learn and perform inference on hand pose data. In the Hand Action paper, LSTM-based methods including [6] and other temporal-dependent methods such as [7] are benchmarked for a hand action recognition task. This project will attempt to re-implement the methods from [5] and [7] using the PyTorch framework with Python3. Simplifications may be introduced if this task proves impractical for the time frame. In this case, the method from [6] is likely easier to implement for hand action recognition.

V. Reading
A multitude of projects and public competitions have tackled problems involving 3D hand pose estimation and action recognition. The listed references [1-7] (and others) will be studied to understand these problems and corresponding state-of-the-art solutions.

VI. Evaluation
The *Hands 2017* competition cites three hand pose estimation error measures as standard:
a. Mean error for all joints for each frame and average across all testing frames.
b. The ratio of joints within a certain error bound.
c. The ratio of frames that have all joints within a certain distance to ground truth annotation.

As for hand action recognition, classification problems are typically evaluated using one or more of the following measures:
a. Accuracy/Error
b. Precision/Recall/F-Measure
c. TPR/FPR/ROC Curve

Measures which are most standard for these problems will identified from an evaluation of the literature and utilized in this project. Plots will be produced to visualize these measures and displayed in the following project submissions.

VII. Timeline
The following timeline is subject to minor changes based on how long each task takes and logistics involving FS Studios and the pico monstar camera. For the midterm progress report, most algorithm development and analysis of the open-source data is expected to be completed.
02/01/2018 – Submit proposal
02/04/2018 – Setup docker image for developing algorithms for the 3D hand image data
02/11/2018 – Develop processing and evaluation framework in python3
02/18/2018 – Program the cited algorithms in PyTorch
02/25/2018 – Test the algorithms on the open-source datasets and benchmark against existing results
02/28/2018 – Submit midterm progress report
03/04/2018 – Acquire and annotate data from the pico monstar camera
03/11/2018 – Test developed algorithms on the pico monstar data
03/19/2018 – Submit presentation video (SCPD student)
03/22/2018 – Submit final project writeup and Git repository link

References

[1] Chang, L., Deng, X., Tan, P., Wang, H., Yang, S., & Zhang, Y. (2017). Hand3D: Hand Pose Estimation using 3D Neural Network. *CoRR, abs/1704.02224*.

[2] Yuan, S., Ye, Q., Stenger, B., Jain, S., & Kim, T. K. (2017, July). Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2605-2613). IEEE.

[3] Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, *33*(5), 169.

[4] Garcia-Hernando, G., Yuan, S., Baek, S., & Kim, T. K. (2017). First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations. *arXiv preprint arXiv:1704.02463*.

[5] Ye, Q., Yuan, S., & Kim, T. K. (2016, October). Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. In *European Conference on Computer Vision* (pp. 346-361). Springer, Cham.

[6] Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016, February). Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks. In *AAAI* (Vol. 2, p. 8).

[7] Zhang, X., Wang, Y., Gou, M., Sznaier, M., & Camps, O. (2016). Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4498-4507).