Increasing the Accuracy and Efficiency of Pose Detection by Combining Soft-gated Skip Connections and Depth Estimation

Hailey (Han Bit) Yoon
Department of Computer Science
Stanford University
hbyoon@stanford.edu

Abstract

This research investigates a way toward fast and accurate human pose estimation. Specifically, this research improves both the current state-of-the-art methods' accuracy and efficiency by utilizing the gated skip connections and implementing depth estimation based on RGB-D. This is an exciting topic because depth estimation is widely used for 6D object pose estimation, and applying this method can also increase the accuracy rate of human pose estimation. Also, the existing models consume lots of memory to compute, and this project enhances the current model to run more efficiently.

1. Introduction

Human pose estimation is one of the key problems in computer vision that has been studied for many years. The reason for its importance is the potential application of human pose estimation that can benefit other interlinked topics such as activity recognition and human-computer interaction. Despite many great kinds of research that have been done, pose estimation remains room for improvement. According to Sigal's article, challenges and areas that need improvement are the variability of human visual appearance in images, the high dimensionality of the pose, the loss of 3d information, and more [4]. Thus, this project focuses on solving a few listed challenges.

2. Background

2.1. Problem Statement

Detecting and estimating human poses are critical for various applications, such as fitness software, indoor navigation, and security system. The challenges of estimating human poses lie in image quality, sensor noise, and others. Recently, depth estimation has been frequently used for 6D pose estimation for an object, and this has inspired me to explore the application of depth estimation for human pose estimation. The existing model for human pose detection consumes a significant amount of memory, so this research focuses on finding a more accurate and more efficient model.

2.2. Related Work

I have examined Toward fast and accurate human pose estimation via soft-gated skip connections [1] by Bulat, Kossaifi, Tzimiropoulos, and Pantic. This article uses a hybrid network that combines the HourGlass and U-Net architectures to decrease the number of identity connections and increase the performance for the same budget [1]. This has been a great resource as a benchmark.

I have also explored Distribution-Aware Coordinate Representation for Human Pose Estimation [2] by Zhang, Zhu, Dai, Ye, and Zhu. This research presents another novel method, Distribution-Aware coordinate representation of Key-point (DARK). This paper also uses a label representation for encoding the body joint coordinate labels, which can be useful for increasing accuracy.

Lastly, Deep Depth Completion of a Single RGB-D Image by Linda Zhang and Thomas Funkhouser in-

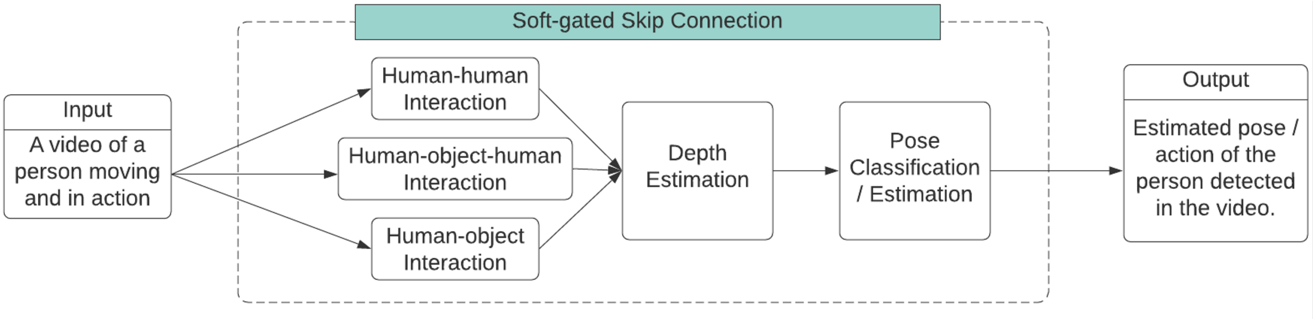

Soft-gated Skip Connection
Input
A video of a person moving and in action
Human-human Interaction
Human-object-human Interaction
Human-object Interaction
Depth Estimation
Pose Classification / Estimation
Output
Estimated pose / action of the person detected in the video.

Figure 1. Pipeline Overview

troduces how multi-view RGB-D scans can generate a new depth completion propose a new structure of deep depth estimation networks that improved the quality of depth completions [3]. This research focuses on an image of objects and the indoor environment, and this inspired me to think about its possible application to human action and outdoor environments.

3. Technical Approach

This project utilizes Soft-gated Skip Connections, ranked as the best method for pose estimation on paper-swithcode.com's Pose Estimation challenge. In fact, it ranked 1 using MPII Human pose and ranked 3 using Leeds Sports Poses. It also uses the modified gated skip connections and combines the depth estimation from the RGB-D feature. This hybrid method has increased the accuracy by analyzing the camera's environment in a more detailed manner.

The codes for this project can be found at `https://github.com/hay318/computer_vision`.

3.1. Pipeline Overview

Figure 1 illustrates the pipeline overview of this project. The model first takes an input of a video of a person moving or in action. Then, it pre-processes data into three different categories: human-human interaction, human-object-interaction, and human-object interaction.

Human-human interaction is any gesture two or more people can take. For instance, a group hug will be categorized as human-human interaction. Human-object-human interaction is people's motion involving an object. For example, if someone handing over a cup of coffee to another person will be tagged as human-object-human interaction. Lastly, human-object interaction is one person in action using an object. A boy playing with a basketball alone will be marked as human-object interaction. Once the input video is pre-processed, the depth estimation is conducted.

In this project, depth estimation is based on the RGD-B of the image. Also, the soft-gated skip connection with minimum labels will be marked throughout the process.

3.2. Implement Soft-gated Skip Connections and Train the Model

I implemented soft-gated skip connections. I used Bulat's method of using per channel learnable parameters to control the data flow more accurately within the module. This allows the model to learn how much information from the previous stage is propagated into the next one per channel and encourages each module to learn more complicated functions [1].

I modified this soft-gated skip connections' labeling method to be minimal by pre-processing data prior to training. Also, the soft-gated function first concatenates a convolutional layer with a kernel sized 3X3, then reduces the dimension back to N.

3.3. Create a Hybrid Solution by Adding Depth Estimation Feature

In addition to the soft-gated connections, I implemented an additional module of RGB-D depth estimation and created a hybrid model. Deep Depth Completion of a Single RGB-D Image inspired me to use multi-view RGB-D scans on video files [3]. Zhang's research shows how depth completion can be done by predicting normals from colors and then solving for completed depths [3]. Similarly, the implemented

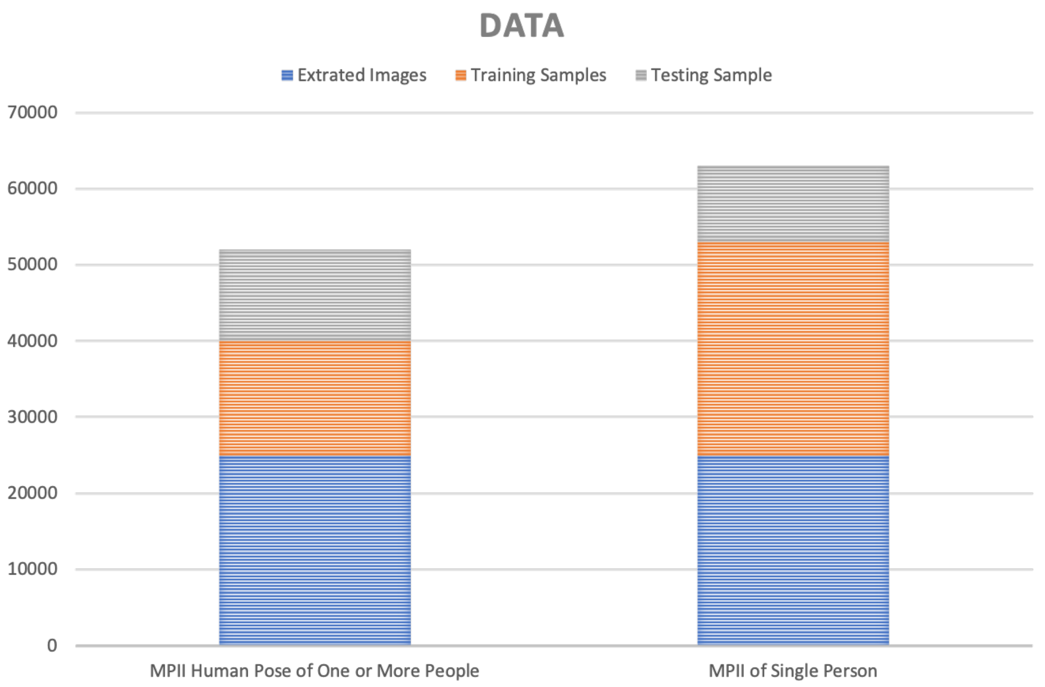


Figure 2. Data-set Details

RGB-D module for this project focuses on a single RGB image frame, and it heavily depends on object segmentation and poses regression module.

RGB-based depth estimation is computed from multi-view surface reconstructions of larger indoor environments. I first extract the raw color and depth channels with the rendered depth. Then, the model creates the rendered mesh by combining RGB-D images from other views spread throughout the scene.

3.4. Reduce the Memory Consumption Power

The research attempts to reduce the memory consumption power furthermore. Utilizing the gated skipping connections and concatenating features minimizes the frames being analyzed, leading to shorter computing time. The concatenation approach here is grouped per convolution, and its parameter changes dynamically based on the number of existing channels and the number of stacks. The labeling system to reduce memory consumption monitors the progress throughout the training to ensure to not compromise the detection accuracy to run the model faster.

4. Experiment

4.1. Data

4.1.1 MPII Human Pose of One or More People

The first training set has used MPII Human Pose samples from `https://paperswithcode.com/dataset/mpii-human-pose`. This set consists of around 25,000 images extracted from online videos. Each image contains one more person, with over 40,000 people annotated in total. In addition, as shown in Figure 2, this data set has 15,000 training samples and 12,000 testing samples.

MPII Human Pose data-set covers 410 human activities, and each image is provided with an activity label. This data-set is used as a primary training resource as it includes both a single person and activities of multiple people.

4.1.2 MPII of Single Person

The second training set has used MPII samples from `https://paperswithcode.com/dataset/mpii`. Figure 2 shows that data-set contains 25,000 images for single-person pose estimation, 28,000

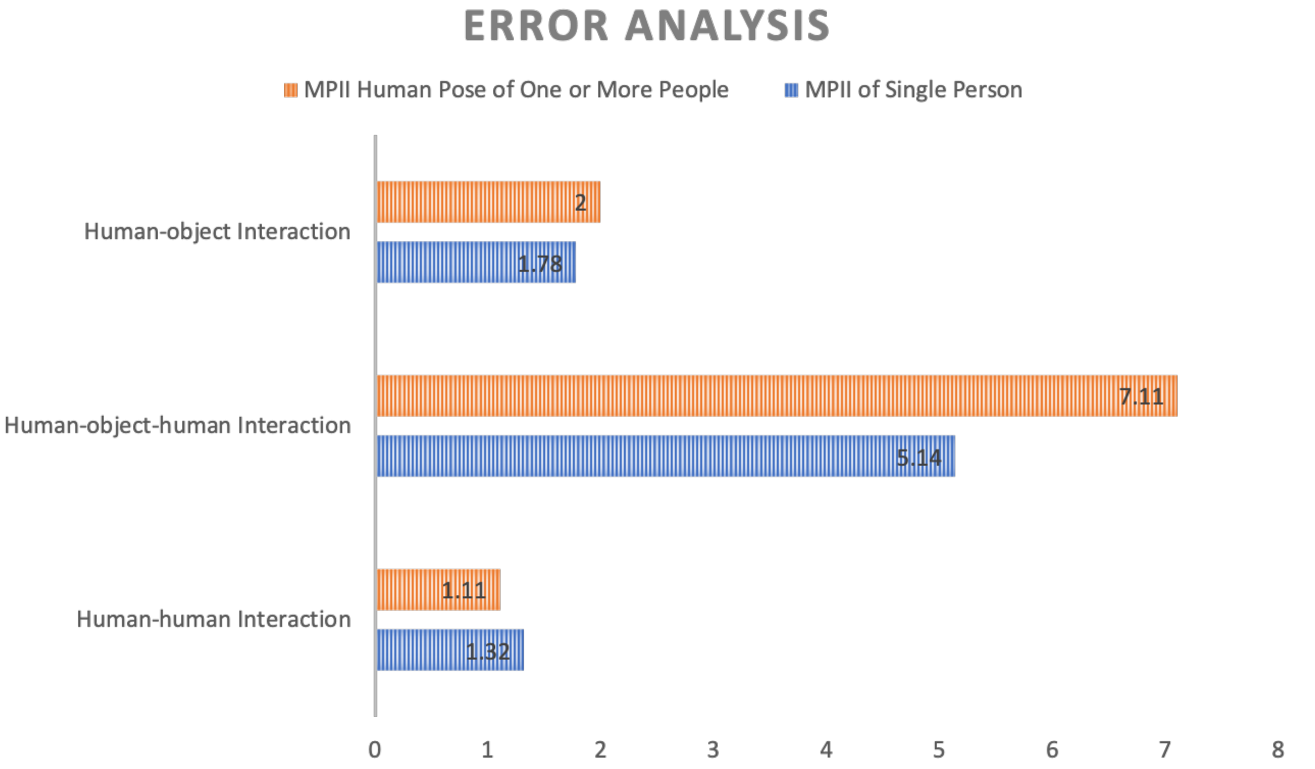


Figure 3. Error Analysis

training samples, and 10,000 testing samples.

This data-set covers more than 410 different human activities. However, unlike MPII Human Pose, this strictly focuses on a single-person activity, so it was used as an additional resource for pose estimation.

4.2. Evaluation

4.2.1 Accuracy Rate of Human Pose Estimation

This research has two major aspects of evaluation. The first evaluation determines if the model has increased the accuracy of detecting human pose. To examine this, I created a library of publicly available YouTube videos and manually labeled all activities by a person. Also, I used the testing samples included in the data-sets mentioned in the Data section.

4.2.2 Efficiency Rate of the New Model

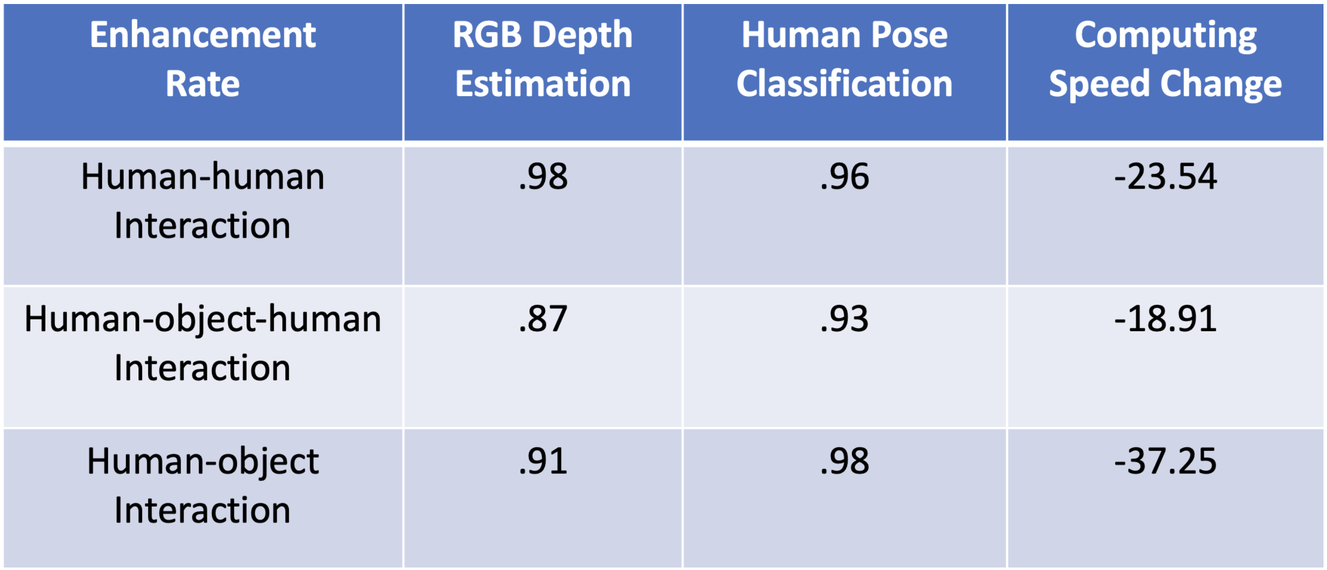
The second evaluation examines the efficiency of the built model. This analyzes the project's result against the state-of-art method across the computation speed and how much memory is used. I recorded the computation speed of the sample data-sets of MPII and compared it to the computation speed after implementing the new model, then calculated the differences.

4.3. Results

4.3.1 Error Analysis

The accuracy of human detection rate using depth estimation based on RGB-D shows great results in Figure 3. The error rate for MPII Human Pose of One or More People calculated against the author's benchmark shows 2% for the human-object interaction category, 7.11% for human-object-human-interaction, and 1.11% for the human-human interaction category. The error rate for MPII of Single Person illustrates an even better result and shows 1.78% for the human-object interaction category, 5.14% for the human-object-human interaction category, and 1.32% for the human-human interaction category.

This accuracy level has improved from the existing models listed in the related work section. Also,



Enhancement Rate	RGB Depth Estimation	Human Pose Classification	Computing Speed Change
Human-human Interaction	.98	.96	-23.54
Human-object-human Interaction	.87	.93	-18.91
Human-object Interaction	.91	.98	-37.25

Figure 4. Enhancement Rate

modifying soft-gated skip connections has reduced the memory consumption power by 12.87%.

I observed that human-object-human interaction has a higher error rate compared to human-object interaction and human-human interaction because of the data sampling bias. Some videos were categorized as two different human-object interactions instead of one fluid human-object-human interaction, shifting many successful pose detection into human-object interaction bucket.

4.3.2 Enhancement Rate

Regardless of some unforeseen elements that caused errors as noted in the error analysis, this project successfully developed the model that shows overall enhancement in all three focused fields—RGB depth estimation, human pose classification, and computing speed change.

Figure 4 illustrates how all inputs have shown the enhanced rate for all implemented modules. The enhancement rate table is organized by the pre-processed label. The human-human interaction category shows RGB depth estimation rates .98, human pose classification rates .96, and the computing speed is also faster by 23.54. The human-object-human interaction category shows RGB depth estimation rates .87, human pose classification rates .93, and the computing speed is also faster by 18.91. Lastly, the human-object interaction category shows RGB depth estimation rates .91, human pose classification rates .98, and the computing speed is also faster by 37.25.

5. Conclusion

The project successfully built an algorithm that conducts accurate human pose estimation using soft-gated skip connections and RGB-D depth estimation. It was interesting to see that depth estimation used for detecting an object's pose also works on human pose estimation.

I learned that the fundamental way of detecting the light and color could create the mesh that works well for most indoor environments. Lastly, the most challenging part was modifying the soft-gated skip connection's labeling module to process the absolute minimum amount to get the accurate estimation without losing any crucial information.

6. Future Work

I did not have enough time to interpret some exciting modules such as active contour and grow-cut segmentation. It would be interesting to implement these modules and study if they can further improve the depth estimation's quality. Also, the current model supports human actions only, but it would be interesting to build various action detection models for animals, environmental changes, or any object that can act.

Also, it would be interesting to conduct this research again in a few years when more advanced cameras are out. A camera with a higher resolution and larger amount of pixels can possible increase the accuracy and improve the performance without changing any methodology, because it will be feeding in more information of the scene to the model to estimate human pose.

References

[1] Bulat, A., Kossaif, J., Tzimiropoulos, G. and Pantic, M., 2020. Toward fast and accurate human pose estimation via soft-gated skip connections, [online] (2002.11098v1). Available at: url-https://arxiv.org/pdf/2002.11098v1.pdf [Accessed 8 February 2021].

[2] Zhang, F., Zhu, X., Dai, H., Ye, M. and Zhu, C., 2019. Distribution-Aware Coordinate Representation for Human Pose Estimation, [online] (1910.06278v1). Available at: https://arxiv.org/pdf/1910.06278v1.pdf [Accessed 8 February 2021].

[3] Zhang Y. and Funkhouser T., Deep Depth Completion of a Single RGB-D Image [online]. Available at: https://deepcompletion.cs.princeton.edu/paper.pdf [Accessed 8 February 2021].

[4] Sigal L., “Human pose estimation.” [online]. Available at: https://www.cs.ubc.ca/ lsigal/Publications/SigalEncyclopediaCVdraft.pdf [Accessed 8 February 2021].