World Scientific Research Journal Volume 6 Issue 11, 2020
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

3D Pose Estimation Using A Combination of CNN and RNN for Replacing the VICON Optoelectronic Motion Capture System

Luqin Chang
Xigongdafuzhong Middle School, Shaanxi, China.

Abstract

Three-Dimensional pose estimation is essential to better understanding the human's movements in different postural tasks. Traditionally, 3D data is captured and processed by the VICON optoelectronic motion capture system. However, the complex operation and high price makes the system uncomfortable to use. In this paper, a new camera-based motion capture system using a novel neural network structure is proposed. We created a new model that combine the Convolutional Neural Network (CNN) and the Recurrent Neural Network. CNN is used as an encoder and RNN is the decoder. After passing the CNN layer, image features are extracted and then fitted to the RNN by guided recurrent unit cell to learn the temporal features. After that, 5 fully connected hidden layers are added to get the prediction for joints. Due to our neural network frame, the image information can be extracted accurately which is the advantage of CNN and the prediction for each frame is depended on the previous frame which is the advantage for RNN to process the sequential data. We experiment our model on the HumanEva dataset, and the result shows that our method can reach the same accurate level as the VICON optoelectronic motion capture system. Therefore, our camera-based motion capture system is a convenient and reliable system to replace the VICON system for 3D pose estimation without losing accuracy.

Keywords

3D pose estimation; RCNN; Motion Capture; VICON system.

1. INTRODUCTION

Pose estimation for human is an important problem and has enjoyed the attention of the Computer Vision community for the past few decades. It is an important step towards understanding people in images and videos. Given an image of a person, 3d pose estimation is the task of producing a 3D pose that matches the spatial position of the person. In order to fix this problem, our neural network has to be invariant to a number of factors, including background scenes, lighting, clothing shape and texture, skin color and image imperfections, among others. This is very difficult.

Usually, people use VICON system to do human's 3D pose estimation. VICON is a marker-based motion capture system and is the most mature tool for tracking the subjects' movements. Precise tracking is essential for providing a ground truth for robot localization experiments. A VICON system provides the highest order of positional precision by optimally selecting and fusing all available data. It is highly accurate and can provide low latency data that is easy to use. Through a special software designed for VICON, the collected data can be integrated into almost any control system.

However, when applying VICON to real experiments, some serious problems occur. First, the VICON system needs a lot of special motion capture cameras. So before starting experiments, researchers need to spend a lot of time to build the support structure and design the locations

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

to set the cameras which is very inconvenient. Besides, in order to capture all motions for a human subject precisely, more than 8 cameras are usually needed. Due to the special camera's price, that is a big cost. What's more, the VICON system's working theory is: The camera of the VICON system emits radiation. The key points on the subject (the important parts in human body like pelvis, trunk, knees and heels... that need to be observed in order to analyze the human motion) will be sticked markers with special materials to reflect the radiation. The camera locates the position of the key points by receiving and analyzing the reflected rays. Before the experiment, researchers need to spend a lot of time cleaning, preparing and sticking the markers which is a waste of time. Also, by applying some devices, the key points might be blocked and cannot be sticked markers.

Considering of all those shortcomings for VICON system, a new monitoring system is designed to use cameras and deep neural networks to replace VICON. Its objective is using video data collected by camera and Deep Neural Network to predict a 3D pose that matches the spatial position of the person. By building, training and tuning the models, the neural network can recognize and predict the key points' position in human subjects.

2. RELATED WORK

Study of human pose estimation started from still image. With traditional methods like pictorial structure [1,2] to newer CNN methods [3]. To improve the performance, Newell et al. [4] proposed Stacked Hourglass Network, which introduced sequential refinement of predictions and the concept of stacked architectures, residual connections and multiscale processing. Replacing the residual unit from the stacked hourglass, Yang et al. [5] proposed a Pyramid Residual Module (PRM). Chu [6] proposed an attention model which is based on conditional random field (CRF). Also, Generative Adversarial Net-works (GANs) are used to improve the capacity to learn structural information.

Distinct from these detection-based methods, regression methods use non-linear function to directly turn the input image into poses in xy coordinates, for example, holistic solution based on cascade regression introduced by Toshev et al. [7] and the Iterative Error feedback proposed by Carreira et al. [8]. Given their advantages of direct pose predictions, regression methods in the literature still give subpar solutions.

As for 3D pose estimation, the situation is even more challenging. Some methods use bottom-up solution, which first locate the body joints, and then predict the 3D poses from it [9]. Sun et al. [10] proposed another method that convert poses into bone figure, which makes it easier to learn and less diverse. Whereas, due to the accumulation of the error between joints, such structural conversion might reduce the precision on joints. To avoid this problem, Pavlakos et al. proposed the Volumetric Stacked Hourglass architecture, which needs precisely estimating volumetric information.

3. METHOD

3.1. Data Pre-Processing

Pytorch is a very efficient tool in deep learning. Before building neural network, we need to build the dataloader for the data. The HumanEva dataset is the most famous dataset in the field of pose estimation. It contains 7 calibrated video sequences (4 grayscale and 3 color) that are synchronized with 3D body poses obtained from a motion capture system (VICON system). In this paper, only one-color camera’s video data is used. For 27 videos in this dataset, we use 23 for training and 4 for testing. The database contains 3 subjects performing 5 common actions which are ‘Gesturing’, ‘Boxing’, ‘Walking’, ‘Catch and Throw’ and ‘Jogging’. (Fig. 1)

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

Figure 1. Five actions covered in HumanEva dataset

The label for one frame is shown in Fig.2, which is a 20*3 array covers the positions for 20 body parts in 3DoF. The unit is mm.

```
'torsoProximal',
'torsoDistal',
'upperLLegProximal',
'upperLLegDistal',
'lowerLLegProximal',
'lowerLLegDistal',
'upperRLegProximal',
'upperRLegDistal',
'lowerRLegProximal',
'lowerRLegDistal',
'upperLArmProximal',
'upperLArmDistal',
'lowerLArmProximal',
'lowerLArmDistal',
'upperRArmProximal',
'upperRArmDistal',
'lowerRArmProximal',
'lowerRArmDistal',
'headProximal',
'headDistal',
```

```
array([[ -944.66303768,     2.43870025,  1286.84916896],
       [ -945.21350987,    36.48908714,   859.04548505],
       [ -915.01125527,   -25.10880093,   847.09909408],
       [ -886.39064899,    -1.65995202,   420.70543547],
       [ -886.36446348,    -1.65039607,   420.65775032],
       [ -911.4451343 ,   330.96742887,   209.07354753],
       [-1006.9175733 ,     5.60620132,   849.66244492],
       [-1094.05281736,   -24.06797595,   459.44579103],
       [-1094.06440087,   -24.07198804,   459.39129134],
       [-1125.46814532,    36.88741116,    93.24302265],
       [ -766.05832919,   -16.73413322,  1259.88723002],
       [ -699.31905491,    87.70186465,  1019.69039721],
       [ -699.31336648,    87.71310998,  1019.66133352],
       [ -808.69657837,  -142.08049911,  1023.48107218],
       [-1073.5888911 ,    66.03676166,  1264.72079248],
       [-1089.52058271,   281.46144677,  1103.41739449],
       [-1089.51084834,   281.43781938,  1103.41947105],
       [-1166.1370691 ,    93.72831299,   943.25955732],
       [ -936.32084077,   -44.3852608 ,  1275.39881709],
       [ -964.97324116,   -76.66605331,  1588.54111222]])
```

Figure 2. An example of data's label

The dataset is created in 2007, due to the technique's limits, not all video frames in the dataset have ground truth, we divide the frames into two kinds, invalid frame and valid frame. Valid frame has ground truth while invalid frame doesn't.

The training dataset's structure is shown in Fig.3. There are 23 videos which are shown as the blue rectangles. Each video has some valid frames which are shown as the red rectangles. The valid frames distributed randomly in the videos which cause a big problem for us. No matter for training or testing, we only need the valid frames.

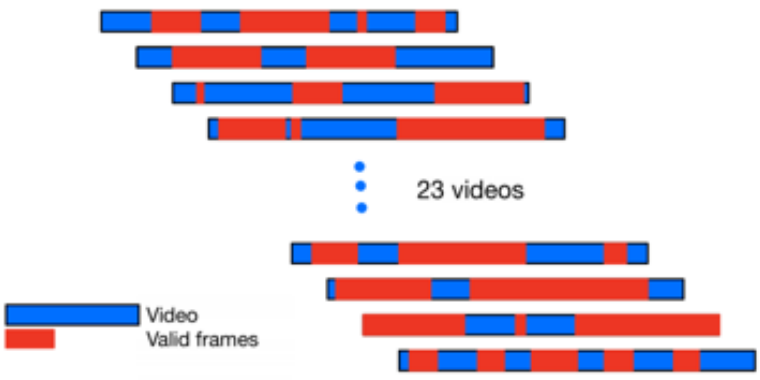

Figure 3. Data Structure

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

A data loader is built based on HumanEva dataset. In the beginning, there are 23 videos. We shuffled the sequence of videos first and then extracted 2 videos as a pair from the videos one by one until all videos are divided into pairs. Batch size is 2 which depends on our computer's memory. For each video, we need to detect and extract all valid frames, almost all of them are continuous but the length of them are quite different between each other.

We randomly choose one continuous valid frame group and compare its length with the sequence-length, which is a number we chose before and shows how many continuous images are there in one sequence. If valid frames' length is smaller than the sequence length, randomly choose the valid frame group again until find a longer one.

Then from the valid frame groups, randomly choose a subgroup of frames as a sequence whose length is the sequence length. Repeat steps 3 to 7 and we can get one mini-batch. After repeating steps 2 to 8 we can get one epoch. In this paper, batch size is 2 and sequence length is 3. Therefore, after these manipulations, one dataloader is built successfully. One epoch covers 11 batches. One batch covers 2 sequences and each sequence includes 3 continuous images with labels.

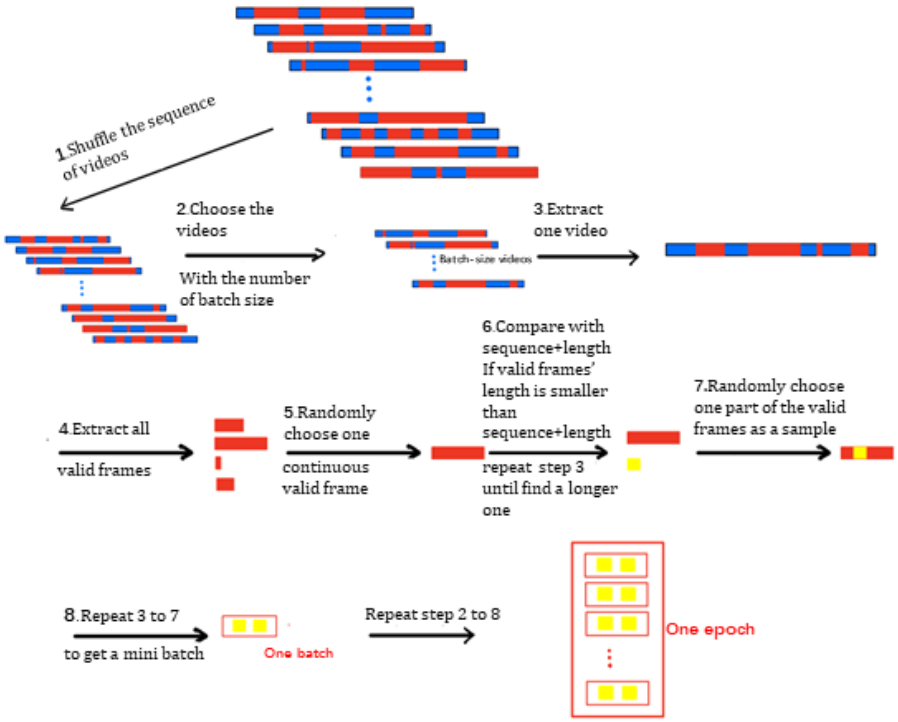

Figure 4. Data Pre-Processing

3.2. Neural Network Structure

Our Neural Network covers 2 parts. The first part is a Convolutional Neural Network (CNN) as an encoder and the other one is an Recurrent Neural Network (RNN) as decoder. In the beginning the pretrained model ResNet50 wat used as backbone to extract features from images. But for the HumanEva dataset, due to the fact that its size is really small, a deeper NN named ResNet152 replaced the ResNet50 which can help get better results. All pretrained models are deleted the last linear layer and we add our own fully connected linear layers to build the CNN. After this step, the features are extracted. Then they are fitted to the RNN by guided recurrent unit cell. After that, 5 fully connected hidden layers are added to get the prediction for joints.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

Due to our NN frame, the prediction for each frame is depended on the previous frame which is the advantage for RNN to process the sequential data.

Table 1. Neural Network's Parameters

Batch size	2
Sequence length	3
Input size	2 * 3 * 3 * image Height * image Width
Output size	2 * 3 * 20 * 3
Loss	Mean Square Error Loss
Optimizer	Adam

After training for 20 epochs the loss plot is like Fig.5. It is evident that the loss decreases quickly in the beginning and converges from about the fourth epoch.

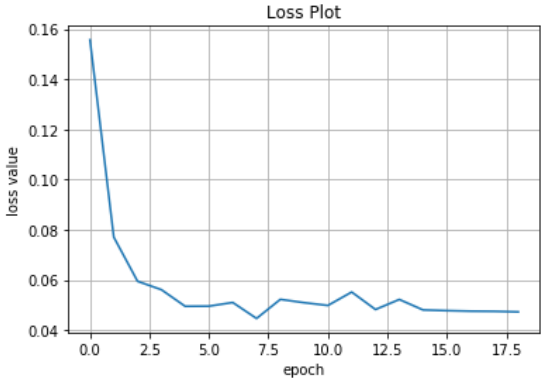

Figure 5. Loss Plot

In order to get one epoch, we need to do a lot of manipulations and it cost a lot of time. We tried a new method. After running one epoch, we shuffled the sequence of the videos and make it as a new epoch for training. After shuffling 9 times we run steps 1~9 (Fig.4) again to get the next epoch. The loss plot is shown in Fig.6. If we set 10 epochs as a cycle, it is evident that the loss decreases sharply in one cycle but increase immediately when beginning a new cycle. This is because for one cycle, the neural network begin overfitting to the data. Therefore, in the end, we returned to the old method to avoid overfitting.

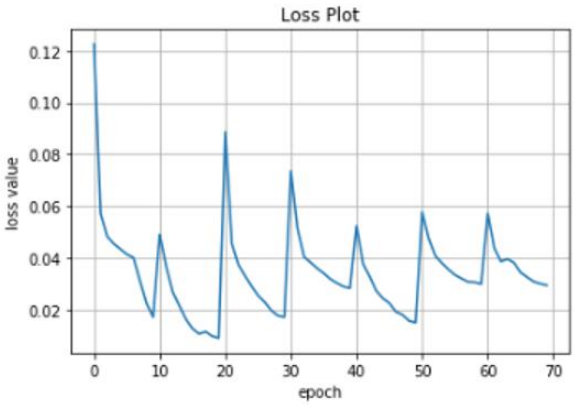

Figure 6. Loss Plot (new method)

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

4. RESULTS AND DISCUSSION

After training the model for 20 epochs, we saved the model's state-dict in every two epochs. After using epoch 5's model and epoch 10's model to predict. We found that there is no big difference in both prediction accuracy. It means after converging, the model did not improve anymore. The accuracy in our project is defined as: (1 - mean error/motion range), the motion range means the biggest moving range of motion, which is got from the whole dataset. The accuracy and mean square error for 20 joints are shown in Fig.7. The left is the accuracy and the right is the mean square error.

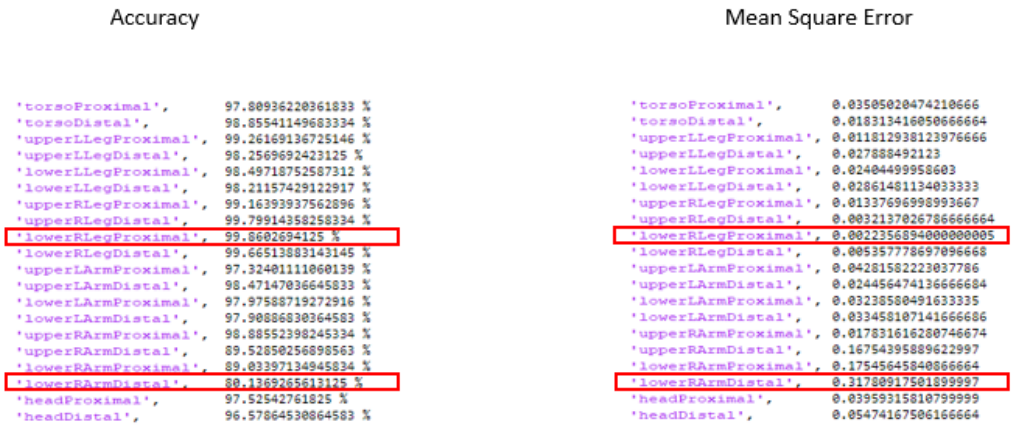

Figure 7. Results

These are for the Boxing action. It is obvious that for boxing, the prediction for the lower right leg proximal is the most accurate, which is about 99% while the lower right arm distal part is the worst, which is about 80%. The average accuracy is about 96% which is very good.

After testing our model in the HumanEva dataset. It is obvious that our model can replace the VICON to detect human's 3D pose. With convolutional neural network, we can do pose estimation for repeated video sequence with high accuracy and it is easier to capture features of pose with larger motion range. However, for VICON system, it can not only detect the positional data, but also the angular data. For our model, we can only detect the joints' positional data. This is one problem we need to solve in the future.

REFERENCES

[1] M.Andriluka, S.Roth, andB.Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2009), pages 1014–1021

[2] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet Conditioned Pictorial Structures. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2013), pages 588–595.

[3] I. Lifshitz, E. Fetaya, and S. Ullman. Human Pose Estimation Using Deep Consensus Voting, (Springer International Publishing, Cham, 2016) pages 246–260.

[4] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. Euro-Conference on Computer Vision (ECCV) (2016), pages 483–499

[5] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In IEEE International Conference on Computer Vision (ICCV) (Oct 2017).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0033

[6] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).

[7] A. Toshev and C. Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In ComputerVision and Pattern Recognition (CVPR), (2014) pages 1653–1660.

[8] Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (June 20 J 16) pages 4733–4742.

[9] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017).

[10] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In IEEE International Conference on Computer Vision (ICCV)(Oct 2017).