Global Pose Refinement using Bidirectional Long-Short Term Memory

Ibrahim Radwan[1], Akshay Asthana [2] and Roland Goecke[1]
[1] The Australian National University, Canberra, Australia
[2] Seeing Machines Ltd., Australia
[3] University of Canberra, Canberra, Australia

Abstract

In This paper, a bi-directional long-short term memory (LSTM) framework is proposed to refine pose estimation and tracking for multiple people. The key idea of our algorithm is to learn the temporal consistencies of the human body shapes between subsequent frames. This helps removing the wrong sudden outliers and improve the general smoothness of the pose tracking. The proposed approach has been evaluated on PoseTrack dataset for both the validation and test subset sequences. The overall detection and tracking results have been improved over the frame-by-frame only baseline detection.

1. INTRODUCTION

Human Pose Estimation is a challenging problem due to its widespread usage in human behaviour analysis, human-computer interaction and affect computing. The difficulties of estimating human poses include inter and intra-occlusion handling [1, 2, 3] and dealing with the highly non-rigidity of the human poses.

The Convolution Neural Network (CNN) has been able to learn and describe the required features of the body parts. However, estimating accurate human pose for multiple people is still hard to achieve. The high non-rigidity of the human body poses and the quick changing in the appearance and the position of the body parts require extra manipulation via imposing the spatial and temporal consistencies to smooth the wrong detections. In this paper, we achieve this goal by investigating temporal relationships between the estimated poses. This leads to smoothing the outliers of the estimated body parts and provides better overall detection.

Recently, Cao *et al.* [4] built a CNN based detection framework to estimate the part location and part affinity and then used a graph approach to estimate the human poses. Their approach did not use the very-important temporal consistency of the body shapes between the subsequent frames. In this paper, we employ these information via adding a bi-directional LSTM on the top of their detection framework. This extra component helps in removing the outliers in the pose detection and provides smoothing tracking results as shown in Fig. 1. The **key contributions** of this paper are:

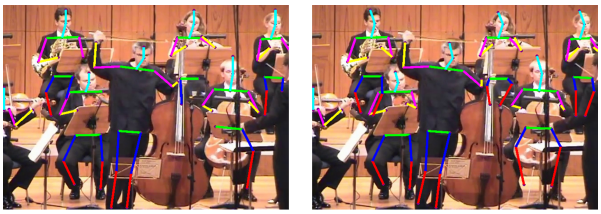

Figure 1. Sample results of the proposed method: The output of applying [4] on frame-by-frame basis (left). The output of adding the B-LSTM on top of the detection phase (Right).

- Encode the temporal consistency between the estimated shapes to improve the quality of the results.
- A pose detection refinement component, which can be plugged on the top of any detection-only method to smooth out the outliers and to provide better tracking outcome.

2. Related Work

In [5], a convolution neural network based model is built to produce the confidence maps of part locations as well as the context of the part with respect to its neighbour parts. Then this approach is extended in [4] via combining the part affinity with the confidence maps to produce a better estimation for the part location. The part affinity helps in associating the adjacent body parts. In this paper, we build upon these two approaches via encoding sequential manner of body shapes in the subsequent frames. This results in improved pose estimation and smooth tracking output.

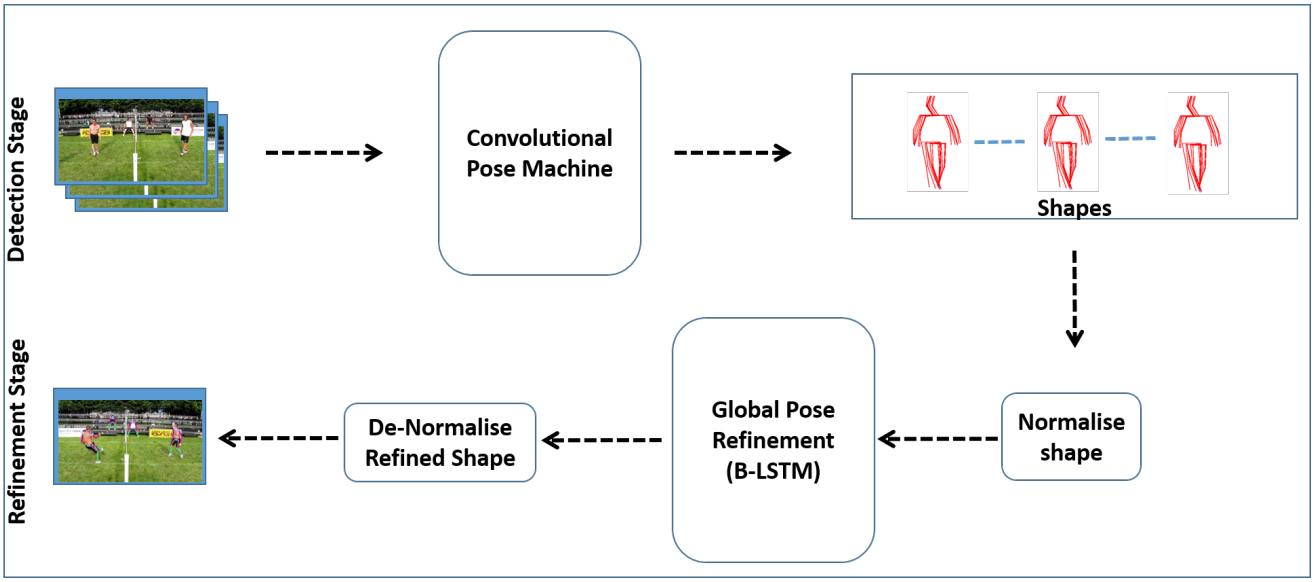

Figure 2. Proposed framework: The input $2n+1$ frames are passed to the detection phase to get the initial estimation of the human poses. Then these poses are normalised and passed to the global pose refinement module. Then the final refinement poses are de-normalised to get the final output.

Also, Pishchulin *et al.* [6] have proposed *DeepCut* approach to estimate the multiple human poses based on employing the Integer Linear Programming (ILP) as an optimisation framework for partitioning and labeling the body part locations. Then, this has been extended in [7] via imposing the spatial relationships between the body parts and learned a CNN-based pairwise features. Again they used the ILP to estimate the part positions.

In [8], the authors built a top-down approach to estimate the human poses. They firstly, employ the Faster-RCNN [9] to detect the scale and location of the persons. Then they use Fully Convolutional Network (FCN) to predict the confidence and offset maps pf the body parts.

In [10], a stacked hourglass network design is proposed to learn the score maps of the body parts as well as the association between the body parts for the same subject.

The above mentioned papers focus on learning the spatial relationships between the body parts. However, in this paper, we propose to build a post-processing components to learn the temporal dependencies between the body pose in successive frames. So that, our proposed method can work on top of any of these approaches to provide better estimation of the body part positions in videos.

3. Proposed Method

As shown in Fig. 2, our proposed framework is organised into two subsequent stages: CNN-based part detection and global pose refinement.

3.1. Body Part Detection

We have employed the recently published work by Cao *et al.* [4] to estimate the body part location for each frame. In their approach, they combined the confidence maps of the body part with the estimation of the part affinity to localise the location of this part. This combination successfully discriminates the body part positions and the associations between the adjacent parts.

3.2. Global Pose Refinement

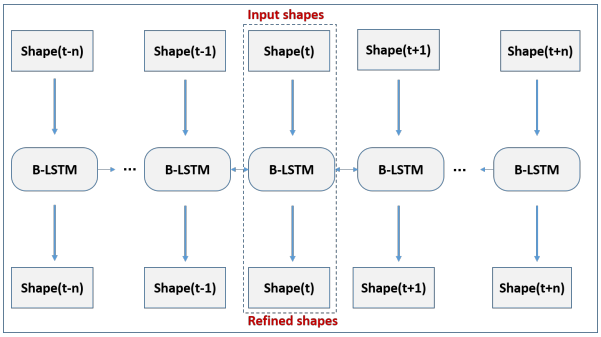

Figure 3. Bidirectional-LSTM many-to-many bidirectional-LSTM to encode the changes in the poses over time. Dashed box surrounds the frame of interest to be refined.

In this section, we describe our proposed approach to refine the detected poses per frame. For each pose $\mathbf{p}_t$ in frame t, We utilize the shape information of the poses from previous and subsequent n frames, *i.e,* $\mathbf{p}_{t-n}$ and $\mathbf{p}_{t+n}$, respectively. We concatenate these shapes subsequently to form

one stream. This stream represents an instance to be processed by the Bidirectional-LSTM. For multiple estimated poses in the video, we keep tracking id for each subject based on the similarity of the estimated shapes.

Firstly, we normalise the shapes in an instance so that the width and height of the shapes will be within unit length. For a given stream, we calculate the coordinates of the bounding box that represents the maximum and the minimum values of the points in the shapes of the stream. Then we centralise each shape into this bounding box and we scale each point in a given shape with the percentage of size of the shape with respect to the size of the bounding box.

Secondly, the normalized shapes in an instance is passed as an input to a bidirectional-LSTM. The LSTM units is capable of encoding the temporal consistency between the subsequent frames. The normalisation step mentioned before is crucial in helping the LSTM units to learn these information as the focus of attention will be in learning the temporal dependencies between the shapes. In this paper, we use all of the points in the estimated pose or shape as a whole item in the stream which lead to learn the global relationships between the shapes in the subsequent frames. This step helps in refining and tuning the estimated poses and pruning any of the outliers that may be happened in the first stage.

In this paper, we employ a many-to-many structure for the LSTM model to learn the mapping between the input normalised shapes, which are extracted from the detection phase and the normalised ground truth shapes. Then we consider only the output of the refined poses for the frame in the middle, *i.e*, frame t. As in Fig. 3, the middle layer is connected together to allow learning the temporal relationships between the subsequent frames. In this paper, we focus on learning the changes in the global pose only, from frame to frame.

Lastly, a de-normalisation step is applied to the shapes which results in the final refined estimation of the poses. This refinement stage improves not only the estimated poses but also the tracking quality.

4. Experiments

We evaluate the performance of our method on the PoseTrack dataset, which has been released as part of the PoseTrack challenge[1]. PoseTrack dataset provides around 464 short video sequences where each sequence is between 50 ad 150 frames. The dataset is split into 250 sequences for training, 50 for validation and 214 for testing. We evaluated our approach on both the validation and testing sequences.

[1] https://posetrack.net

4.1. Evaluation Protocols

The first metric which is used per-frame and for multiple pose estimation is the Average Precision (AP) as in [6]. Every prediction on the frame is assigned to the ground truth and the highest PCK value [11] is selected. In our experiment we report the results with Object Keypoint Similarity (OKS$= 0.5$). The second metric is Multiple Object Tracking (MOT) [12] where tracking identification is defined for each object and distance between the predicted and ground truth is computed. For both metrics, we have used the evaluation server that is provided with the PoseTrack challenge [2].

4.2. Discussion

Table 1. Per-frame multiple pose estimation on PoseTrack validation and test set. A comparison between the proposed approach (detection + B-LSTM) and detection only.

Validation set								
Method	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
Detection only	49.2	69.1	63.7	52.3	56.7	53.3	47.3	55.5
Detection + LSTM	51.2	70.4	64.0	52.5	56.7	53.6	47.6	56.2
Test set								
Detection + LSTM	51.9	69.4	62.3	51.3	56.7	53.0	47.7	55.8

The B-LSTM model is trained on the training sequences, where the input data are the initial frame-by-frame shapes that are estimated in the detection phase. The target shapes are the ground truth. We have investigated different designs for the LSTM structure with various *Dense* and *TimeDistributed* layers for inputs and outputs. We fixed the activation functions to be linear for all of the design and used $L2$ regularisation with a fixed value (0.0001). The model that gives the best results in this experiments includes two many-to-many bi-directional LSTM layers.

While the LSTM model is trained on the training sequences, we kept the validation set a side for testing the algorithm before and after adding the recurrent component. The results are reported on Tables 1 and 2 for both the AP and MOT metrics respectively. The proposed method shows slight improvement over the baseline as the LSTM has to learn the global changes in the poses between frames. Although, the high non-rigidity of the used data and the lack of

[2] https://github.com/leonid-pishchulin/poseval

Table 2. Multiple Object Tracking (MOT) results on PoseTrack validation and test set. A comparison between the proposed approach (detection + B-LSTM) and detection only.

Validation set											
	MOTA	MOTA	MOTA	MOTA	MOTA	MOTA	MOTA	MOTA	MOTP	Prec	Rec
Method	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total	Total	Total	Total
Detection only	-39.3	-8.1	-41.7	-65.2	-23.3	-44.2	-67.4	-41.2	nan	50.1	69.1
Detection + LSTM	-5.1	35.4	-3.0	-32.5	17.1	-7.8	-34.7	-4.1	nan	50.8	69.9
Test set											
Detection + LSTM	1.9	35.5	-14.3	-49.0	1.8	-29.8	-59.0	-14.9	43.0	47.7	70.0

the training data (only 250 short sequences), the proposed method still provides a bit of improvement, which we believe that with more training sequences, the results would be better. In the below parts of the tables, we also put the results of applying the algorithm on the testing sequences.

We focus the attention on predicting the missing and the body parts with low detection scores as we know that the LSTM is powerful in predicting based on the history and the future data. This is why, we designed our algorithm to work only on the body parts, which their score is below a threshold (0.5 in our experiments). The score has been calculated as the average of scores of pixels around a detected joint for a body part. The score is the combination of the confidence map values and the part affinity values. Based on this score, we determine either to refine the part or just keep it as detected.

In Fig. 4, we show qualitative comparison between the proposed algorithm (right column) and without adding the LSTM layer (left column). The results show the ability of the LSTM to rectify the missing body parts. Please note the the existence and absence of the legs and arms of the persons in the images.

5. CONCLUSIONS

In This paper, we have added a recurrent layer in terms of Bi-directional LSTM on top of a CNN based framework for multiple pose estimation. The B-LSTM has shown its capability of encoding the global dependencies and learning the temporal information between frames. This has been shown in rectifying the positions of missing and low-scores body parts. We believe that the LSTM has the capability to learn the spatial dependencies between the local parts in the same frame, however this will be investigated in a future work.

ACKNOWLEDGMENT

We applaud the great effort of the PoseTrack dataset team for collecting and preparing the data, which helped us in pursuing this work. Also, we acknowledge the support of Nvidia company for granting us a Titan-XP GPU device that helped us to implement our experiments.

References

[1] I. Radwan, A. Dhall, and R. Goecke, "Occlusion-aware human pose estimation with mixtures of subtrees," *CoRR*, vol. abs/1512.01055, 2015. 1

[2] I. Radwan, A. Dhall, and R. Goecke, "Monocular image 3d human pose estimation under self-occlusion," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 1888–1895, IEEE, 2013. 1

[3] I. Radwan, A. Dhall, J. Joshi, and R. Goecke, "Regression based pose estimation with automatic occlusion detection and rectification," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pp. 121–127, IEEE, 2012. 1

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Real-time multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017. 1, 2, 5

[5] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016. 1

[6] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4929–4937, 2016. 2, 3

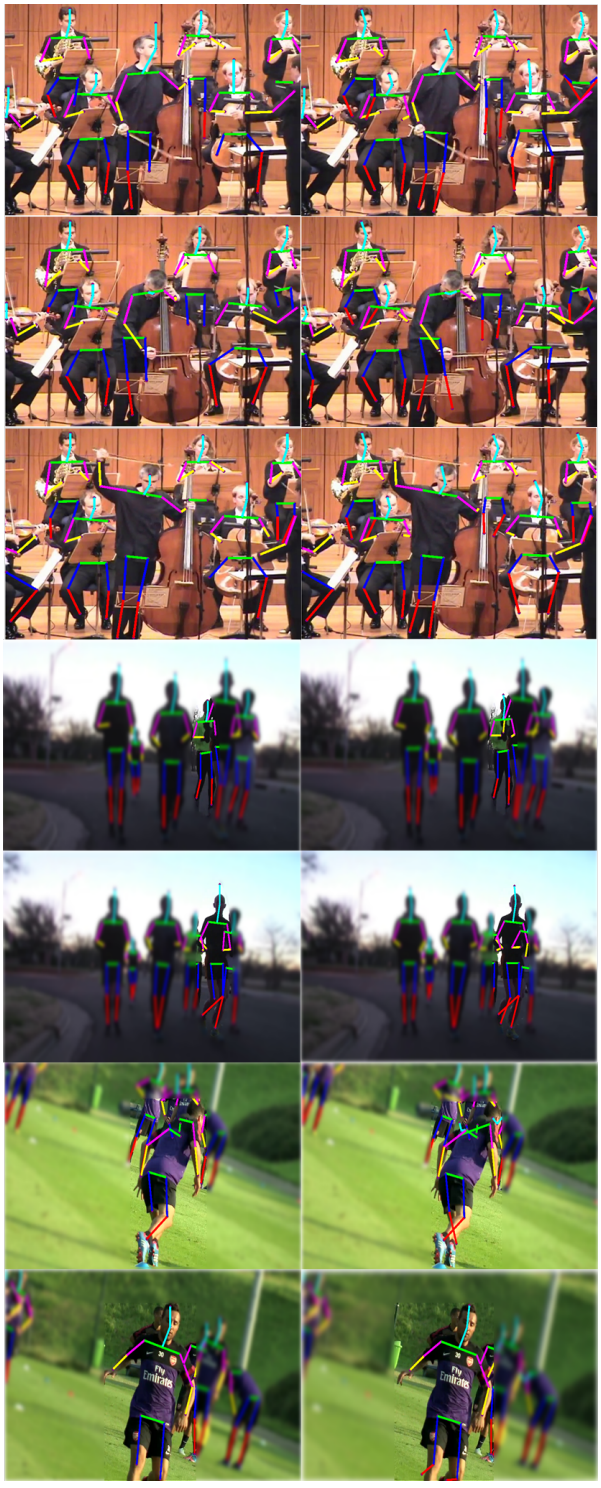

Figure 4. Visual comparison to show the effectiveness of using the B-LSTM on top of the detection (right) and without using it as in [4] (left).

[7] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, pp. 34–50, Springer, 2016. 2

[8] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy, "Towards accurate multi-person pose estimation in the wild," *CoRR*, vol. abs/1701.01779, 2017. 2

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[10] A. Newell, K. Yang, and J. Deng, *Stacked Hourglass Networks for Human Pose Estimation*, pp. 483–499. 2016. 2

[11] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3

[12] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, 2008. 3