ESTIMATION OF MISSING HUMAN BODY PARTS VIA BIDIRECTIONAL LSTM

Ibrahim Radwan[1], Akshay Asthana[2], Hafsa Ismail[3], Byron Keating[4] and Roland Goecke[3]
[1] The Australian National University, Canberra, Australia
[2] Seeing Machines Ltd., Canberra, Australia
[3] Human-Centred Technology Research Centre, University of Canberra, Canberra, Australia
[4] Queensland University of Technology, Brisbane, Australia

***Abstract*— In this paper, a bi-directional long-short term memory (LSTM) based approach is proposed for the estimation of missing body parts in a human pose estimation context. Accurate human pose estimation is often a key component for accurate human action and activity recognition. The key idea of our algorithm is to learn the temporal consistencies of the human body poses between previous and subsequent frames. This helps in estimating missing body parts and improves the general smoothness of the pose detection results. The approach acts as a post-processing step after the application of any off-the-shelf body part detector and has been evaluated on the PoseTrack dataset for both validation and testing sequences. The results show consistent improvement in the detection across all body parts.**

I. INTRODUCTION

Human pose estimation is an active area of research due to its widespread usage in human behaviour analysis, human-computer interaction and affective computing. The challenges in estimating human body parts include drastic variation in the appearance of individual body parts (due to environment and clothing etc.), inter- and intra-occlusion handling [12], [10], [11] and dealing with the high non-rigidity of human poses. The ability of estimating missing body parts due to occlusion or lack of appearance features leads to more accurate human body action and activity recognition as well as effective gesture analysis and understanding.

Convolution Neural Networks (CNN) have been able to address these challenges to a significant extent but estimating accurate human pose for multiple people in a video sequence consistently is still an open problem. In particular, the issue of missing body parts detection due to high non-rigidity of the human body poses, overlapping body parts and quick changes in the appearance and the position of the body parts require additional constraints via imposing spatial and temporal consistencies. This can provide smoother and complete detections on a frame-by-frame basis. In this paper, we achieve this goal by investigating temporal relationships between the estimated poses and provide a mechanism to estimate the missing body parts, which in turn 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. While this framework provides exceptional results on a frame-by-frame basis, the approach did not use the very-important temporal consistency of the body poses between the subsequent frames.

There are two generic ways to enforce this temporal consistency. Firstly, we can train the detection network end-to-end to detect body pose in the entire sequence (or a sub-sequence) simultaneously. Secondly, by treating the output of the detection network on a frame-by-frame basis and training a separate network, which takes as input the results of the detection network for the entire sequence (or a sub-sequence) and generates temporally consistent results. Given that the problem of estimating the body pose for the entire sequence simultaneously is highly non-linear, owing to the number of possible combinations of types of appearances and movements a body in a sequence can exhibit, the limiting factor for the first approach is the requirement of a large enough dataset of sequences, annotated with all body parts for every frame, to cover a decent number of possible combinations for the network to generalize. On the other hand, for the second approach, we can remove the appearance component from the problem by treating the body part estimation provided by the detection network as input to the second network. The second network can now address the problem of enforcing temporal consistency taking the body pose of the entire sequence (or sub-sequence) as input. In this paper, we follow the second approach and enforce the temporal consistency in estimating the body pose by adding a bi-directional LSTM on top of the detection framework [4]. We show that this method can be used effectively to estimate missing parts (Fig. 1).

II. RELATED WORK

In [3], a CNN-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] by combining the part affinity with the confidence maps to produce a better estimation for the part locations. The part affinity helps in associating the adjacent body parts.

In this paper, we build upon these two approaches via encoding temporal information of the body poses present on the previous and subsequent frames. This results in a more complete pose estimation output. Also, Pishchulin *et al.* [1] proposed the *DeepCut* approach to estimate multiple human poses based on employing Integer Linear Programming (ILP)

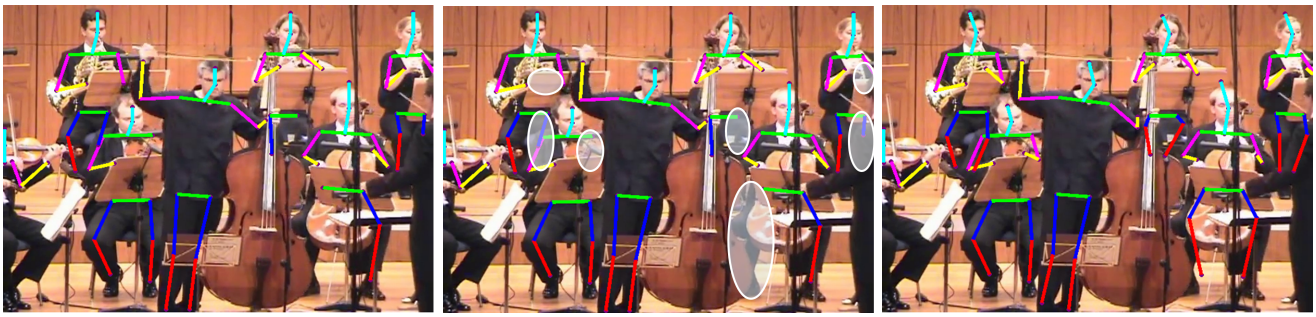

Fig. 1. Sample results of the proposed method: Applying [4] on a frame-by-frame basis (left). Highlighted missing body parts (Middle). After applying B-LSTM on top of the detection step, the estimated missing body parts are included (Right).

as an optimisation framework for partitioning and labeling the body part locations. This approach was extended in [2] by imposing spatial relationships between the body parts and learning a CNN-based pairwise features. Again, ILP was used to estimate the part positions. In [5], a top-down approach to estimate human poses was built. Firstly, the Faster-RCNN [6] was employed to detect the scale and location of the body. Then, a Fully Convolutional Network (FCN) was used to predict the confidence and offset maps of the body parts. In [7], a stacked hourglass network design was 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 literature focuses on learning the spatial relationships between the body parts. However, in this paper, we propose to build post-processing components to learn the temporal dependencies between the body pose. Our proposed method can work on top of any of the above mentioned approaches and estimate the missing body parts to provide better and more complete pose estimation results.

III. Proposed Method

The proposed framework consists of two stages: CNN-based part detection and Bidirectional LSTM (B-LSTM) based missing part estimation (Fig. 2).

A. Body Part Detection

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

B. Estimation of Missing Parts via B-LSTM

In this section, we describe our proposed approach to estimate the missing body parts. For each pose $\mathbf{p}_t$ in frame t, we utilise 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 poses 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 detected subject based on the similarity of the estimated poses.

For a given stream, we calculate the coordinates of the bounding box that represents the maximum and minimum values of the points in the poses of the stream. Each pose is then centred into this bounding box and normalised such that the width and height of the poses are within unit length. Next, the normalised pose in an instance is passed as an input to a Bidirectional-LSTM, which is capable of encoding the temporal consistency between the previous and subsequent frames.

In this paper, we employ a many-to-many structure for the B-LSTM model to learn the mapping between the normalised input poses, which are noise induced ground-truth poses, and the normalised ground truth poses. To be specific, we induce three types of variation to generate the noisy input stream from the ground-truth poses in that we randomly simulate missing parts, missing frames in the stream and add noise to the individual parts in the stream. Overall, we simulate 100 noisy input streams per ground truth input stream where up to 50% parts are randomly removed from each stream to simulate missing body parts, up to 50% frames are randomly removed from each stream to simulate missing frames, and 10% of streams are induced with small translation noise (2% noise in X- and Y-direction) to each body part independently to simulate noisy detections. In short, the goal is to enable the network to learn to extract the underlying ground-truth stream from a noisy input stream.

Once the network has been trained, the output from the body part detection step is passed through the network to generate the refined body poses, which include the missing body parts locations as well. We consider only the output of the refined poses for the frame in the middle, *i.e* frame t as the final output (Fig. 3). Lastly, a de-normalisation step is applied to the output pose, which results in the final refined estimation of the body pose, which includes the locations of the missing body parts in image co-ordinates. The location of these missing body parts is then merged with the results from the detector to yield the final complete body pose. Overall, the proposed approach acts as a post-processing step, which helps in estimating the missing body parts to yield a more complete body tracking result for the stream.

The proposed method is designed to work offline with recorded videos, where both of the estimations of the previous and future frames are available. However, the proposed

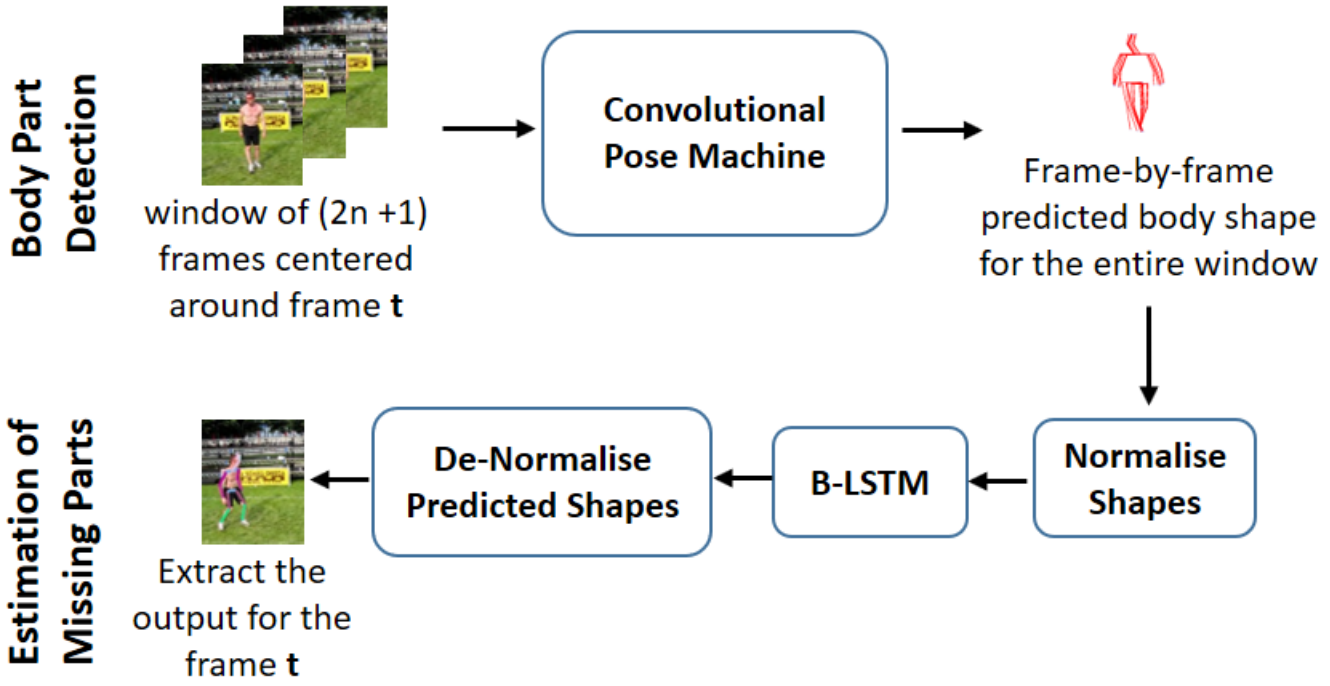

Fig. 2. The input of $2n+1$ frames, centred around frame t, is passed to the body part detection phase for the initial estimation of the body poses. Then, the predicted poses are normalised and passed to B-LSTM, which predicts the refined poses for each frame. These poses are de-normalised and the final pose for frame t is obtained.

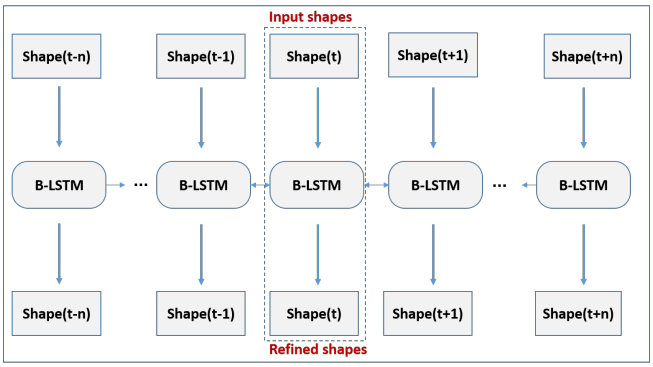

Fig. 3. The many-to-many bidirectional-LSTM encodes the changes in the poses over time. The dashed box surrounds the frame of interest to be refined.

method can be extended to a uni-directional LSTM working only with the previous estimations and, hence, online and real-time analysis can be achieved. Moreover, the proposed method can also be used with a fixed (and minor) latency of n number of frames, where n can be configured as per the application.

IV. Experiments

We evaluate the performance of the proposed method on the PoseTrack dataset, which has been released as part of the PoseTrack challenge[1]. The PoseTrack dataset provides 464 short video sequences where each sequence is between 50 ad 150 frames long. The dataset is split into 250 sequences for training, 50 for validation and 214 for testing. This dataset contains diverse challenging scenarios such as highly occluded parts, scenes with a different number of persons and with changing illuminations. We evaluated our approach on both the validation and testing sequences. However, note that the ground-truths for the testing sequences have not been released by the challenge organisers, therefore, we provide a detailed breakdown of the missing body parts only for the validation set.

A. Evaluation Protocols

The metric, which is used per-frame and for multiple pose estimation, is the Average Precision (AP) as in [1]. Every prediction on the frame is assigned to the ground truth and the highest PCK value [8] is selected. In the experiments, we report the results with Object Keypoint Similarity (OKS$=$ 0.5). For the purpose of reporting the results presented in this paper, we use the evaluation server that is provided with the PoseTrack challenge [2].

B. Discussion

The B-LSTM model is trained on the training sequences, where the input data consists of the noise induced ground truth shapes and the output is the actual ground truth shape (Section III-B). We have investigated different designs for the B-LSTM structure with multiple B-LSTM layers, different recurrent network configurations (many-to-one and many-to-many), dense layers and time-distributed dense layers for

[1] https://posetrack.net

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

Fig. 4. Qualitative results of the proposed method. For each triplet of an example image, we show the results of applying [4] on a frame-by-frame basis (left), B-LSTM predicted missing body parts (middle), and combined result of applying B-LSTM on top of the detection step, including the estimated missing body parts (right).

TABLE I

RESULTS ON POSETRACK VALIDATION AND TESTING SET.

Parts Name	Validation set						Testing set
	GT parts	Parts detected by Cao *et al.* [4]	AP for Cao *et al.* [4] (%)	Parts not detected Cao *et al.* [4]	AP for B-LSTM (%)	Overall AP Cao *et al.* [4] + B-LSTM (%)	Overall AP Cao *et al.* [4] + B-LSTM (%)
Head	14591	10414	49.2	4177	56.18	51.2	51.9
Shoulder	30058	28235	69.1	1823	90.53	70.4	69.4
Elbow	25765	21715	63.7	4050	65.6	64	62.3
Wrist	24760	17650	52.3	7110	52.99	52.5	51.3
Hip	29376	26539	56.3	2837	60.44	56.7	56.7
Knee	27217	21341	53.3	5876	54.68	53.6	53.0
Ankle	24642	16785	47.3	7857	48.24	47.6	47.7
Total	**176444**	**142679**	**55.5**	**33765**	**59.15**	**56.2**	**55.8**

inputs and outputs. We fixed the activation functions to be rectified linear units (*ReLU*) for all designs and used $L2$-regularisation with a fixed value (0.0001) for each layer. The standard Mean Absolute Error (MAE) between the predicted and the ground truth is used as the loss function. Overall, we did not see any significant difference in the performance by adding multiple dense or time-distributed dense layers and a simple network design, which includes two B-LSTM layers with 512 units each followed by a time-distributed output layer, enabling the many-to-many LSTM configuration, gave the best result.

Since we do not have access to the ground truths of the testing set (as mentioned above), we kept the validation set aside for testing the algorithm before and after adding the recurrent component. For training the network, we used 230 training sequences as the training set and the remaining 20 training sequences as the validation set. The results are reported in Table I using the AP metric. We focus the

attention on predicting the missing body parts to leverage on the ability of the B-LSTM to use temporal information from previous and future frames. For the combined detection and B-LSTM results, we replace the location of those missing body parts with the results from B-LSTM.

The proposed method shows improvement for all the body parts over the baseline (i.e. using only the detector [4]). The overall AP of 55.5% improved after the application of B-LSTM to 56.2%. However, note that the overall improvement does not clearly reflect the utility of the proposed method since we only predict the missing body parts. Overall, roughly 20% of the parts were reported missing by [4]. The B-LSTM obtains an overall AP of 59.15% on these missing body parts. We also show the results on the PoseTrack testing set where we obtain an overall AP of 55.8%. In Fig. 4, we show a qualitative comparison between the proposed framework with and without adding the B-LSTM module. The results show the ability of the B-LSTM to estimate the missing body parts accurately.

V. CONCLUSIONS

In this paper, we have proposed a B-LSTM based framework to estimate the locations of missing body parts in human pose estimation. The B-LSTM has shown a promising capability of learning the temporal information between the frames. The framework acts as a post-processing module and can be used with any off-the-shelf body part detector to predict the locations of the missing body parts in a sequence. In the future, we plan to extend this method to incorporate CNN features from the detectors in the learning phase of the B-LSTM and perform an end-to-end refinement over the entire sequence. However, we believe that the training dataset of 250 short sequences (each sequence have only 30 frames ground-truth annotation available) is not sufficient to model the highly non-linear subspace of the temporal motion of the body parts and CNN features. Therefore, we plan to extend the dataset by annotating all frames in the training dataset sequence and incorporate other datasets, which have annotated body sequences.

Acknowledgment: This research was funded in part by the Australian Research Council (ARC) Linkage Project grant LP160100910. Moreover, we acknowledge the support of NVIDIA for granting us a Titan-XP GPU device that helped us to implement our experiments.

REFERENCES

[1] Pishchulin, Leonid and Insafutdinov, Eldar and Tang, Siyu and Andres, Bjoern and Andriluka, Mykhaylo and Gehler, Peter V and Schiele, Bernt, "Deepcut: Joint subset partition and labeling for multi person pose estimation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp 4929-4937.

[2] Insafutdinov, Eldar and Pishchulin, Leonid and Andres, Bjoern and Andriluka, Mykhaylo and Schiele, Bernt, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model", European Conference on Computer Vision (ECCV), 2016, pp 34-50.

[3] Shih-En Wei and Varun Ramakrishna and Takeo Kanade and Yaser Sheikh, "Convolutional pose machines", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] Zhe Cao and Tomas Simon and Shih-En Wei and Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[5] George Papandreou and Tyler Zhu and Nori Kanazawa and Alexander Toshev and Jonathan Tompson and Chris Bregler and Kevin P. Murphy, "Towards Accurate Multi-person Pose Estimation in the Wild, CoRR, 2017.

[6] Shaoqing Ren and Kaiming He and Ross Girshick and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Advances in Neural Information Processing Systems (NIPS), 2015.

[7] Alejandro Newell and Kaiyu Yang and Jia Deng, "Stacked Hourglass Networks for Human Pose Estimation", European Conference on Computer Vision (ECCV), 2016, pp 483-499.

[8] Andriluka, Mykhaylo and Pishchulin, Leonid and Gehler, Peter and Schiele, Bernt, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis", IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[9] Bernardin, Keni and Stiefelhagen, Rainer, "Evaluating multiple object tracking performance: the CLEAR MOT metrics", EURASIP Journal on Image and Video Processing, 2008.

[10] Radwan, Ibrahim and Dhall, Abhinav and Goecke, Roland, "Monocular Image 3D Human Pose Estimation under Self-Occlusion", IEEE International Conference on Computer Vision (ICCV), 2013, pp 1888-1895

[11] Radwan, Ibrahim and Dhall, Abhinav and Joshi, Jyoti and Goecke, Roland, "Regression based pose estimation with automatic occlusion detection and rectification" IEEE International Conference on Multimedia and Expo (ICME), 2012, pp 121-127

[12] Radwan, Ibrahim and Dhall, Abhinav and Goecke, Roland, "Occlusion-Aware Human Pose Estimation with Mixtures of Sub-Trees", CoRR, abs/1512.01055, 2015