# A Bottom-Up Approach for Pig Skeleton Extraction Using RGB Data

Akif Quddus Khan, Salman Khan, Mohib Ullah(✉), and Faouzi Alaya Cheikh

Norwegian University of Science and Technology, 2815 Gjøvik, Norway
mohib.ullah@ntnu.no

**Abstract.** Animal behavior analysis is a crucial task for the industrial farming. In an indoor farm setting, extracting Key joints of animals is essential for tracking the animal for a longer period of time. In this paper, we proposed a deep network that exploits transfer learning to train the network for the pig skeleton extraction in an end to end fashion. The backbone of the architecture is based on an hourglass stacked dense-net. In order to train the network, keyframes are selected from the test data using K-mean sampler. In total, 9 Keypoints are annotated that gives a brief detailed behavior analysis in the farm setting. Extensive experiments are conducted and the quantitative results show that the network has the potential of increasing the tracking performance by a substantial margin.

**Keywords:** Pig · Behavior analysis · Hourglass · Stacked dense-net · K-mean sampler

## 1 Introduction

Automatic behavior analysis of different animal species is one of the most important tasks in computer vision. Due to variety of applications in the human social world like sports player analysis [1], anomaly detection [2], action recognition [3], crowd counting [4], and crowd behavior [5,6], humans have been the main focus of research. However, due to the growing demands of food supplies, vision-based behavior analysis tools are pervasive in the farming industry and demands for cheaper and systematic solutions are on the rise. From the algorithmic point of view, other than the characterization of the problem, algorithm design for humans and the farm animals are similar. Essentially, behavior analysis is a high-level computer vision task and consists of feature extraction, 3D geometry analysis, and recognition, to name a few. As far as the input data is concerned, it could be obtained through smart sensors (Radio-frequency identification [7], gyroscope [8], GPS [9]). Depending on the precision of measurements, such sensors give acceptable results but using such sensors has many drawbacks. For example, in most cases, it is required to remove the sensor from the animal to collect the data. Such a process is exhausting for the animals and laborious for the human operator. Compared to this, a video-based automated behavior

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 54–61, 2020.
https://doi.org/10.1007/978-3-030-51935-3_6

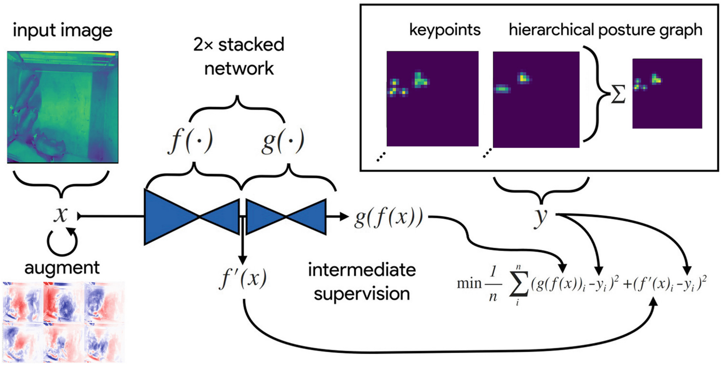

**Fig. 1.** Detailed illustration of model training process

analysis offers a non-invasive solution. Due to cheaper hardware, it is not only convenient for the animals but also cost-effective for the industry. Automatic behavior analysis and visual surveillance [10–12] has been used for the security of public places (airports, shopping malls, subways, etc.) and turned into a mature field of computer vision.

In this regard, Hu et al. [13] proposed a recurrent neural network named MASK-RNN for the instance level video segmentation. The network exploits the temporal information from the long video segment in the form of optimal flow and perform binary segmentation for each object class. Ullah et al. [14] extracted low level global and Keypoint features from video segments to train a neural network in a supervised fashion. The trained network classifies different human actions like walking, jogging, running, boxing, waving and clapping. Inspired from the human social interaction, a hybrid social influence model has been proposed in [15] that mainly focused on the motion segmentation of the moving entities in the scene. Lin et al. [16] proposed a features pyramid network that extract features at different level of a hierarchical pyramid and could potentially benefit several segmentation [13,17,18], detection [19,20], and classification [21,22] frameworks. Additionally, in the field of cybersecurity [23,24], such techniques are very beneficial. By addressing the problem of scale variability for object detection, Khan et al. [20] proposed a dimension invariant convolution neural network (DCNN) that compliment the performance of RCNN [19] but many other state-of-the-art object detectors [4,25] could take advantage of it. Inspired by the success of deep features, [26] proposed a two-stream deep convolutional network where the first stream focused on the spatial features while the second stream exploit the temporal feature for the video classification. The open-source deep framework named OpenPose proposed by Cao et al. [27] focuses on the detection of Keypoints of the human body rather than the detection of the whole body. Detection of Keypoints has potential applications in the pose estimation and consequently behavior analysis. Their architecture consists of two convolutional neural networks were the first network extract features and gives the location of the main joints of the human body in the form of a heat map. While the second network is responsible for associating the corresponding body joints. For the feature extraction, they used the classical VGG architecture. The frameworks like OpenPose are very helpful in skeleton extraction of the human body and potentially, it could be used in tracking framework. For example, the Bayesian framework proposed in [28] works as a Keypoint tracker, where any Keypoints like the position of head [29], or neck or any other body organ can be used to do tracking for longer time. Such Keypoints can be obtained from a variety of human pose estimation algorithms. For example, Sun et al. [30] proposed a parallel multi-resolution subnetworks for the human pose estimation. The idea of a parallel network helps to preserver high resolution and yield high quality features maps that results in better spatial Keypoints locations. Essentially, in such a setting, the detection module is replaced by [27,30]. In this regard, a global optimization approach like [31] could be helpful for accurate tracking in an offline setting. By focusing only on pose estimation of humans, Fang et al. [32] proposed a top-down approach where first the humans are detected in the form of bounding boxes and later, the joints and Keypoints are extracted through a regional multi-person pose estimation framework. Such a framework is helpful in not only in the localization and tracking of tracking in the scene, but also getting the pose information of all the targets sequentially. For a robust appearance model that could differentiate between different targets, a sparse coded deep features framework was proposed in [33] that accelerate the extraction of deep features from different layers of a pre-trained convolution neural network. The framework helps handle the bottleneck phenomenon of appearance modeling in the tracking pipeline. Alexander et al. [34] used transfer learning [35] and fine-tuned ResNet to detect 22 joints of horse for the pose estimation. They used the data collected from 30 horses for the within domain and out of domain testing. The work by Mathis et al. [36] analyzed the behavior of mice during the experimental tasks of tail-tracking, and reach & pull of joystick tasks. They also analyze the behavior of drosophila while it lays eggs. The classical way of inferring behavior is to perform segmentation and tracking [37] first, and based on the temporal evolution of the trajectories, perform behavior analysis. However, approaches like [38] can be used to directly infer predefined actions and behaviors from the visual data. In addition to the visual data, physiological signals [39,40], and acoustic signals [41] can be used to identify different emotional states and behavioral traits in farm animals.

Compared to the existing methods, our proposed framework is focused on the extraction of the key joints of the pig in an indoor setting. The visual data is obtained from a head-mounted Microsoft Kinect sensors. Our proposed framework is inspired by [42] where a fully-convolutional stacked hourglass-shaped network is proposed that converts the image into a 16-channel space representing detection maps. For the part detection, the thresholds are set from 0.10 to 0.90. These thresholds are used while evaluating the recall, precision, and F-measure metrics for both the vector matching and euclidean matching results. Such an analysis provides a detailed overview of the trade-offs between precision and recall while maintaining an optimal detection threshold. The loss function, the optimizer, and the training details are also given in Sect. 3. The qualitative results are mentioned in Sect. 4 and the remarks are given in Sect. 5 which concludes the paper.

The rest of the paper is organized in the following order. In Sect. 2 the proposed method is briefly explained including the Keypoints used in the experiment, the data filtration and annotation, and the augmentation. Model architecture along with the loss function, the optimizer, and the training details are elaborated in Sect. 3. The qualitative results are given in Sect. 4 and the remarks are given in Sect. 5 which concludes the paper.

## 2 Proposed Approach

The block diagram of the network is given in Fig. 1. It mainly consists of two encoder-decoder stacked back to back. The convolution neural network used in each encoder-decoder network is based on the dense net. The network takes the input as the visual frame. To train the model, we annotate the data by first converting videos into individual frames, and then annotating each frame separately by specifying the important key points on the animal's body. After sufficient training, the model returns a $9 \times 3$ matrix for each frame, where each row corresponds to one keeping, the first two columns specify the x and y coordinates of the detected point, and the third column contains the confidence score of the model. After obtaining the x and y coordinates for each frame, we visualize these key points on each frame and stitch all the individual frames into a single video. A total of nine (Nose, Head, Neck, Right Foreleg, Left Foreleg, Right Hind leg, Left Hind Leg, Tail base, and tail tip) key points is being focused for each pig.

### 2.1 Data Filtration

The given RGB data consists of three sets with three pigs, six pigs, and ten pigs. Each dataset has 2880 images. To get better and more accurate results, a larger dataset was required. However, K-Mean clustering is applied to each dataset for selecting the most informative frames. As a result, only 280 images are extracted from the larger dataset. The count is approximately 10% of the size of the original dataset. Data annotator developed by Jake Graving is used to annotate the dataset. It provides a simple graphical user interface that reads key points data from the CSV file and saves the data in .h5 format once the annotation is completed. DeepPoseKit works with augmenters from the imgaug package. We used spatial augmentations with axis flipping and affine transforms.

## 3 Model Architecture

The proposed framework is based on an hourglass densenet which is an efficient multi-scale deep-learning model. The architecture consists of an encoder and decoder where dense nets are stacked in sequential order. Densenet is a densely Connected Convolutional Networks [43]. DenseNet can be seen as the next generation of convolutional neural networks that are capable of increasing the depth of the model with every decreasing the number of parameters.

### 3.1 Loss Function and Optimizer

Mathematically, the loss function is:

$$L(x, y) = \frac{1}{n} \sum_{i}^{n} ((g(f(x))_i - y_i)^2 + (f'(x) - y_i)^2) \tag{1}$$

where x is the input sample and y corresponds to the network prediction. We used the callback function using *ReduceLROnPlateau.* ReduceLROnPlateau automatically reduces the learning rate of the optimizer when the validation loss stops improving. This helps the model to reach a better optimum at the end of training. While training a model on three pigs' data, first a test was run for data generators. Creating a TrainingGenerator from the DataGenerator for training the model with annotated data is an important factor. The TrainingGenerator uses the DataGenerator to load image-keypoints pairs and then applies the augmentation and draws the confidence maps for training the model. The validationSplit argument defines how many training examples to use for validation during training. If a dataset is small (such as initial annotations for active learning), we can set this to validationSplit = 0, which will just use the training set for model fitting. However, when using callbacks, we made sure to set monitor = "loss" instead of monitor = "valloss". To make sure the Training Reduce learning rate parameters saves useless resource utilization and model overfeeding. For this particular reason, the parameter is set to reduce the learning rate by 0.2 if the loss does not improve after 20 iterations. Another parameter that is used to prevent resource exploitation is Early Stopping. Patience is set 100 iterations which means training would stop automatically if the loss does not improve after 100 iterations. Training started at a loss of 220, after running 400 iterations, the loss stopped showing improvement at 4.5. In the test case, when given the same video from which dataset was generated, very accurate results are produced.

## 4 Experiments

The proposed framework is implemented in Python with the support of Keras backend by Tensorflow. The processing is performed on Nvidia P-100 with 32 GB RAM. The qualitative results are shown in Fig. 2. It can be seen that the keypoints are successfully extracted from the pig joints. However, sometimes, when the pigs are very close to each other, the extracted keypoints are associated with the wrong animal. It is simply because the network is trained on very limited data. Training the network with more data to help improve the association of keypoints.

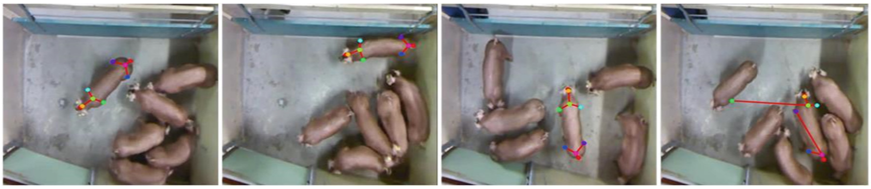

**Fig. 2.** Qualitative results of the proposed method.

## 5 Conclusion

We proposed a deep network for animal skeletonization based on only RGB data. The network exploits use varies data augmentation and transfer learning to fine-tune the parameters. The backbone of the network is based on an hourglass stacked dense-net. In order to train the network, keyframes are selected from the test data using K-mean sampler. In total, 9 Keypoints are annotated that gives a brief detailed behavior analysis in the farm setting. Experiments are conducted on pig data and the quantitative results that training the network with only 280 frames yields promising results.

## References

1. Khan, S.D., et al. Disam: density independent and scale aware model for crowd counting and localization. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 4474–4478. IEEE (2019)
2. Ullah, H., Altamimi, A.B., Uzair, M., Ullah, M.: Anomalous entities detection and localization in pedestrian flows. Neurocomputing **290**, 74–86 (2018)
3. Yang, J., Shi, Z., Ziyan, W.: Vision-based action recognition of construction workers using dense trajectories. Adv. Eng. Inform. **30**(3), 327–336 (2016)
4. Khan, S.D., et al.: Person head detection based deep model for people counting in sports videos. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE (2019)
5. Ullah, M., Ullah, H., Conci, N., De Natale, F.G.B.: Crowd behavior identification. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1195–1199. IEEE (2016)
6. Ullah, H., Ullah, M., Conci, N.: Dominant motion analysis in regular and irregular crowd scenes. In: Park, H.S., Salah, A.A., Lee, Y.J., Morency, L.-P., Sheikh, Y., Cucchiara, R. (eds.) HBU 2014. LNCS, vol. 8749, pp. 62–72. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11839-0_6
7. Maselyne, J., et al.: Measuring the drinking behaviour of individual pigs housed in group using radio frequency identification (RFID). Animal **10**(9), 1557–1566 (2016)
8. Ullah, M., Ullah, H., Khan, S.D., Cheikh, F.A.: Stacked LSTM network for human activity recognition using smartphone data. In: 2019 8th European Workshop on Visual Information Processing (EUVIP), pp. 175–180. IEEE (2019)
9. Pray, I.W., et al.: GPS tracking of free-ranging pigs to evaluate ring strategies for the control of cysticercosis/taeniasis in Peru. PLoS Negl. Trop. Dis. **10**(4), e0004591 (2016)

10. Alreshidi, A., Ullah, M.: Facial emotion recognition using hybrid features. In: Informatics, vol. 7, p. 6. Multidisciplinary Digital Publishing Institute (2020)
11. Chen, J., Li, K., Deng, Q., Li, K., Philip, S.Y.: Distributed deep learning model for intelligent video surveillance systems with edge computing. IEEE Trans. Ind. Inform. (2019)
12. Ullah, H.: Crowd motion analysis: segmentation, anomaly detection, and behavior classification. Ph.D. thesis, University of Trento (2015)
13. Hu, Y.-T., Huang, J.-B., Schwing, A.: MaskRNN: instance level video object segmentation. In: Advances in Neural Information Processing Systems, pp. 325–334 (2017)
14. Ullah, M., Ullah, H., Alseadonn, I.M.: Human action recognition in videos using stable features (2017)
15. Ullah, H., Ullah, M., Uzair, M.: A hybrid social influence model for pedestrian motion segmentation. Neural Comput. Appl. **31**, 7317–7333 (2018). https://doi.org/10.1007/s00521-018-3527-9
16. Lin, T.-Y., et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
17. Ullah, H., Uzair, M., Ullah, M., Khan, A., Ahmad, A., Khan, W.: Density independent hydrodynamics model for crowd coherency detection. Neurocomputing **242**, 28–39 (2017)
18. Ullah, M., Mohammed, A., Alaya Cheikh, F.: PedNet: a spatio-temporal deep convolutional neural network for pedestrian segmentation. J. Imaging **4**(9), 107 (2018)
19. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
20. Khan, S., et al.: Dimension invariant model for human head detection. In: 2019 8th European Workshop on Visual Information Processing (EUVIP), pp. 99–104. IEEE (2019)
21. Wei, Y., Sun, X., Yang, K., Rui, Y., Yao, H.: Hierarchical semantic image matching using cnn feature pyramid. Comput. Vis. Image Underst. **169**, 40–51 (2018)
22. Ullah, M., Ullah, H., Cheikh, F.A.: Single shot appearance model (SSAM) for multi-target tracking. Electron. Imaging **2019**(7), 466-1 (2019)
23. Yamin, M.M., Katt, B.: Modeling attack and defense scenarios for cyber security exercises. In: 5th Interdisciplinary Cyber Research Conference 2019, p. 7 (2019)
24. Yamiun, M.M., Katt, B., Gkioulos, V.: Detecting windows based exploit chains by means of event correlation and process monitoring. In: Arai, K., Bhatia, R. (eds.) FICC 2019. LNNS, vol. 70, pp. 1079–1094. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-12385-7_73
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
26. Ullah, H., et al.: Two stream model for crowd video classification. In: 2019 8th European Workshop on Visual Information Processing (EUVIP), pp. 93–98. IEEE (2019)
27. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: real-time multi-person 2D pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
28. Ullah, M., Cheikh, F.A., Imran, A.S.: Hog based real-time multi-target tracking in Bayesian framework. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 416–422. IEEE (2016)

29. Ullah, M., Mahmud, M., Ullah, H., Ahmad, K., Imran, A.S., Cheikh, F.A.: Head-based tracking. In: IS&T International Symposium on Electronic Imaging 2020: Intelligent Robotics and Industrial Applications using Computer Vision proceedings, San Francisco, USA 2020. Society for Imaging Science and Technology. https://doi.org/10.2352/ISSN.2470-1173.2020.6.IRIACV-074
30. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019)
31. Ullah, M., Cheikh, F.A.: A directed sparse graphical model for multi-target tracking. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1816–1823 (2018)
32. Fang, H.-S., Xie, S., Tai, Y.-W., Lu, C.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343 (2017)
33. Ullah, M., Mohammed, A.K., Cheikh, F.A., Wang, Z.: A hierarchical feature model for multi-target tracking. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2612–2616. IEEE (2017)
34. Mathis, A., Yüksekgönül, M., Rogers, B., Bethge, M., Mathis, M.W.: Pretraining boosts out-of-domain robustness for pose estimation (2019)
35. Ullah, M., Kedir, M.A., Cheikh, F.A.: Hand-crafted vs deep features: a quantitative study of pedestrian appearance model. In: 2018 Colour and Visual Computing Symposium (CVCS), pp. 1–6. IEEE (2018)
36. Mathis, A., et al.: Markerless tracking of user-defined features with deep learning. arXiv preprint arXiv:1804.03142 (2018)
37. Ullah, M., Cheikh, F.A.: Deep feature based end-to-end transportation network for multi-target tracking. In: IEEE International Conference on Image Processing (ICIP), pp. 3738–3742 (2018)
38. Nasirahmadi, A., Edwards, S.A., Sturm, B.: Implementation of machine vision for detecting behaviour of cattle and pigs. Livestock Sci. **202**, 25–38 (2017)
39. Kanwal, S., et al.: An image based prediction model for sleep stage identification. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1366–1370. IEEE (2019)
40. Atlan, L.S., Margulies, S.S.: Frequency-dependent changes in resting state electroencephalogram functional networks after traumatic brain injury in piglets. J. Neurotrauma **36**, 2558–2578 (2019)
41. da Cordeiro, A.F.S., et al.: Use of vocalisation to identify sex, age, and distress in pig production. Biosyst. Eng. **173**, 57–63 (2018)
42. Psota, E.T., Mittek, M., Pérez, L.C., Schmidt, T., Mote, B.: Multi-pig part detection and association with a fully-convolutional network. Sensors **19**(4), 852 (2019)
43. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)