Proceedings of the ASME 2018
Dynamic Systems and Control Conference
DSCC2018
September 30-October 3, 2018, Atlanta, Georgia, USA

DSCC2018-9007

PARALLEL DEEP LEARNING ENSEMBLES FOR HUMAN POSE ESTIMATION

Hailin Ren
Dept. of Mechanical Engineering
Virginia Tech,
Blacksburg, VA, USA

Anil Kumar
Dept. of Mechanical Engineering
Virginia Tech,
Blacksburg, VA, USA

Xinran Wang
Dept. of Mechanical Engineering
Virginia Tech,
Blacksburg, VA, USA

Pinhas Ben-Tzvi*
Dept. of Mechanical Engineering
Virginia Tech,
Blacksburg, VA, USA

ABSTRACT

This paper presents an efficient method to detect human pose with monocular color imagery using a parallel architecture based on deep neural network. The network presented in this approach consists of two sequentially connected stages of 13 parallel CNN ensembles, where each ensemble is trained to detect one specific kind of linkage of the human skeleton structure. After detecting all skeleton linkages, a voting score-based post-processing algorithm assembles the individual linkages to form a complete human structure. This algorithm exploits human structural heuristics while assembling skeleton links and searches only for adjacent link pairs around the expected common joint area. The use of structural heuristics in the presented approach heavily simplifies the post-processing computations. Furthermore, the parallel architecture of the presented network enables mutually independent computing nodes to be efficiently deployed on parallel computing devices such as GPUs for computationally efficient training. The proposed network has been trained and tested on the COCO 2017 person-keypoints dataset and delivers pose estimation performance matching state-of-art networks. The parallel ensembles architecture improves its adaptability in applications aimed at identifying only specific body parts while saving computational resources.

Keywords: Pose Estimation, Convolutional Neural Networks (CNN), Linkage-based Approach, Parallel CNN Architecture.

INTRODUCTION

As recent progress in computational capabilities enable machines to come into the real world from a lab setting, it becomes important to understand and study nearby human activity to address safety concerns. Human pose estimation is already an active research problem for machine perception systems in self-driving cars, search and rescue systems, automated surveillance and other Human-Robot Interaction (HRI) applications [1]. Accurate and efficient human pose estimation is critical in achieving high-level tasks such as pedestrian avoidance, automated robotic lifting and moving victims for search and rescue applications, human behavior recognition, etc.

Based on the application and the availability of sensing modality, the input data could be either 2D images [2–6], 3D point clouds [7,8], one single frame or a sequence of frames (motion-tracking) [9]. In recent decades, researchers have focused on model-based algorithms, which deploy finely tuned feature extractors such as SIFT [5] and HoG [10] along with different human models such as Pictorial Structures [11] and Active Shape Models [6]. In more recent years, as artificial intelligence has become significantly popular with the HRI researchers due to the advancements in computing technologies, exploration of neural networks for human pose estimation has also picked up the pace. On the same track, Toshev and Szegedy [2] utilized sequentially connected convolution layers and fully connected layers to build one deep neural network for high precision pose estimation. Li et al. proposed a pose-joint repressor with body-part detector using a single deep neural network [4]. In a slightly different approach, Tompson et al. proposed a hybrid architecture using both convolutional neural networks and Markov Random Field [3]. While other researchers explored human pose estimation in still-imagery, the computer vision group from UC Berkeley looked into the use of Recurrent-CNNs for both pose estimation and gait/action recognition in video input [12,13]. Recognition of the pose of a single person [5,6,14–18] in an image sets up the foundation for pose estimation of multiple persons [9,12,19–22].

One straightforward approach for multi-person pose-estimation is to apply a person detector on the input image, and then for each person detector proposal, apply a single person

*Corresponding author – bentzvi@vt.edu

pose estimation method. This approach is called a top-down method [21,22] and suffers from early commitment, meaning that there is no chance to detect a person that is not proposed by the person detector, such as Faster RCNN [23]. The computational cost of this approach is proportional to the number of detector proposals from one image. Another strategy for multi-person pose estimation is the bottom-up method [19]. In these models, the network detects the 'body parts' of the person visible in the input imagery first and then assembles them to multiple individuals according to a given policy. This kind of approach solves the early commitment problem but suffers from a computational cost problem, since associating different parts to individuals is an NP-hard problem [20].

One such example of the bottom-up approach builds upon a deep convolutional network to detect joints. A higher-level spatial model is then used to constrain joint inter-connectivity and generate the global pose [3]. Insafutdinov et al. used a strong detector to detect person's joints and assemble those joints using image-conditioned pairwise terms [19].

One hybrid approach incorporating both the bottom-up and top-down methods has been developed. Sheng et al. proposed a method by which one bottom-up detector is used to detect joints and one top-down human detector is used to rule out the bottom-up false alarms resulting in significantly improved tracking accuracy [24].

Cao et al. proposed the use of a joint heat map with part affinity fields to estimate 2D human poses by using multi-stage, sequentially connected CNN branches [25]. In their approach, each branch was responsible for predicting one specific body linkage in terms of joint heat maps and a 2D vector field (Part Affinity Fields (PAF)) representing the direction shape of the linkage mask.

This paper builds on the work by Cao et al. [25] and presents a deep neural network architecture for multi-person pose estimation using parallel CNN ensembles. Similar to previous work, the proposed neural network assumes that the body skeleton consists of linkages connected to joints. However, this paper focuses on exploring the effect of parallelization of the CNN nodes on the performance of the bottom-up human pose estimation. The network is trained to estimate the location and orientation of each link and the location of joints. Greedy parsing is then used to assemble the linkages to human individuals. The algorithm has been trained and tested using the COCO 2017 person keypoint dataset.

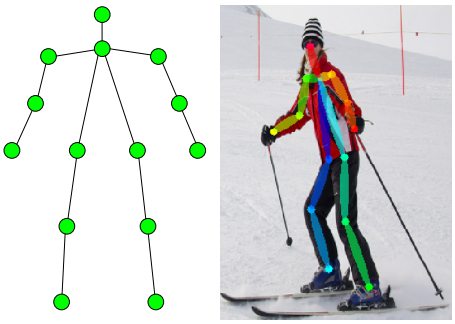

Figure 1. Proposed 13-link human skeleton model.

PROPOSED NETWORK ARCHITECTURE

The proposed system models the human pose using a 13-link, 14 joint skeleton model as shown in Fig. 1. The method presented in this paper consists of two tasks: (1) body part detection, and (2) pose regression. The body parts are denoted by a set $P_n = \{P_1, P_2, ..., P_N\} \in \mathbb{R}^{w \times h \times 3}$, where $P_n \in \mathbb{R}^{w \times h \times 3}$ and N is a set to 13 representing the number of parts detected by the proposed neural network. Each part P_n consists 2D vector fields of the link L_n and a confidence map of two associated joints J_n, where $\mathrm{L}_n \in \mathbb{R}^{h \times w \times 2}$ and $\mathrm{J}_n \in \mathbb{R}^{h \times w}$. The two associated joints are denoted as high joint J_n^h and low joint J_n^l separately. J_n^h denotes the joint that is connected with a former body part while J_n^l denotes the joint that is connected with a the later body part; more specifically, $\mathrm{J}_n^h \in \mathbb{R}_{>0}^{h \times w}$, $\mathrm{J}_n^l \in \mathbb{R}_{<0}^{h \times w}$. For example, 'right shoulder - right elbow' body part consists of two associated joints J_n, right shoulder joint and right elbow joint and one link L_n that connects these two joints. In this body part, right shoulder joint is defined as high joint J_n^h while right elbow joint is the low joint J_n^l.

In the body part detection task, the proposed neural network takes one single 3-channel color image with size (h×w×3) as the input and produces the 2D location and orientation tensor of the size (h'×w'×39) for each body part of persons visible in the input image. The pose regression process then assembles the body parts for all individuals in the input image using greedy inference.

Overall Architecture Design:

As shown in Fig. 2, the proposed neural network architecture can be divided into 3 stages: preprocessing stage, prediction stage 1, and prediction stage 2. In the preprocessing

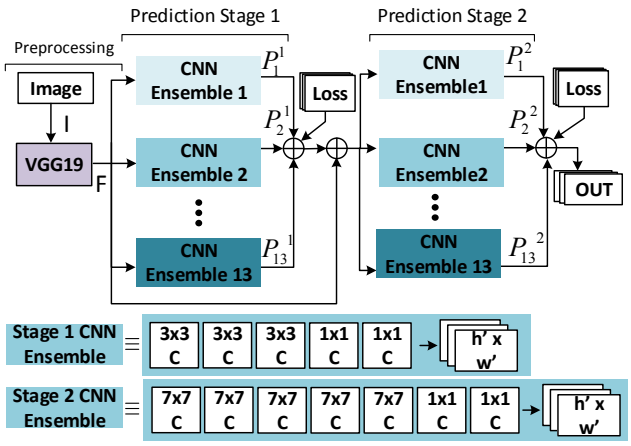

Figure 2. Architecture of the 2-stage 13-parallel CNN ensemble network.

stage, the color input image, I , is fed to a pre-trained VGG19 network [26] to obtain feature map F. In each predicting stage, there exist 13 CNN ensembles to predict the link fields and joint confidence maps for each of the 13 skeletal model linkages independently. The first predicting stage takes the feature map as inputs and produces 2D location and orientation tensor, which can be denoted as $P_n^1 = \rho_n^1(F), n \in \{1,...,13\}$. The outputs of the first prediction stage merge with the feature maps F to generate one single tensor of size h'×w'×167 which then serves as the input for the second prediction stage. The second prediction stage's output can be denoted as $P_n^2 = \rho_n^2(P^1, F), n \in \{1,...,13\}$. The independent architecture of each branch is aimed at achieving an independent prediction behavior for each linkage. The merging of the stage one output with the original feature maps to be fed to stage two is aimed at providing a reference to stage two for refining the linkage predictions of each body part. The outputs of each branch present the position and orientation of a body part as a tensor of size h'×w'×3, whose last dimension represents the number of channels. The first and second channels present the X and Y component of the PAF of the link, respectively. The third channel is the heat map of the two associated joints.

Supervision is provided at the end of each stage. To train the network to detect all the thirteen body linkages, the loss function for each part has been defined by incorporating the confidence of the associated joints and the PAF for each linkage. In multi-person images, the net loss function of the network consists of contribution from individual labelled poses. Based on the policy applied by [25], weighted functions are used to compute the total loss, f, as follows:

$$f = \sum_i {}^iW \, {}^if \tag{1}$$

where ${}^iW$ is the binary mask indicator. It is zero when the annotation to the i^{th} person is missing. Therefore, ${}^iW$ helps avoid penalizing the neural network when the ground truth is missing in the dataset. The total loss function of the i^{th} person, ${}^if$, which includes the loss function of all the body parts is

$${}^if = {}^if_J + {}^if_L = \sum_p ({}^if^p_{JH} + {}^if^p_{JL} + {}^if^p_L) \tag{2}$$

where ${}^if_J$ and ${}^if_L$ are the loss functions for the joint's confidence map and link PAF, respectively. Due to the different properties of the two associated joints, ${}^if^p_J$ is further divided into ${}^if^p_{JH}$ and ${}^if^p_{JL}$ for the high and low joints of the p^{th} body part. The loss functions for the joints and the links are presented as follows:

$$\begin{gathered} {}^if^p_{JH} = \left\| J^p_{H,i} - {}^*J^p_{H,i} \right\|_2^2, \; {}^if^p_{JL} = \left\| J^p_{L,i} - {}^*J^p_{L,i} \right\|_2^2 \\ {}^if^p_L = \left\| L^p_i - {}^*L^p_i \right\|_2^2 \end{gathered} \tag{3}$$

where, ${}^*J^p_H$, ${}^*J^p_L$, and ${}^*L^p$ are the loss functions defined for the high joints, low joints, and links of the p^{th} body, respectively.

Body Part Presentation:

To evaluate the loss function, a three channel-ground truth for each body part was generated for each image pixel using the annotated 2D keypoints from the COCO dataset. The ground truth label is generated for each image pixel wherever any body part element is visible. The location and orientation of each link is represented using PAF while the joint location is expressed using a bipolar Gaussian joint confidence map. The high joint has a positive confidence map while the low joint uses a negative confidence map to distinguish between each other. Let $\boldsymbol{x}^p_{jh,i} \in \mathbb{R}^2, \boldsymbol{x}^p_{jl,i} \in \mathbb{R}^2$ be the ground truth of the high joint and the low joint of the p^{th} body part for the i^{th} person. The values of $s^p_{jh,i}$, $s^p_{jl,i}$ for this joint at any pixel location $\boldsymbol{x} \in \mathbb{R}^2$ are calculated as follows:

$$s^p_{jh,i} = \exp\left(-\frac{\left\| \boldsymbol{x} - \boldsymbol{x}^p_{jh,i} \right\|_2^2}{\sigma^2} \right), s^p_{jl,i} = -\exp\left(-\frac{\left\| \boldsymbol{x} - \boldsymbol{x}^p_{jl,i} \right\|_2^2}{\sigma^2} \right) \tag{4}$$

where the standard deviation, σ, controls the spread of the peaks. The aggregation of the high joint's and the low joint's ground truth are the maximum absolute values of the individuals, with the assumption that no high joint for any linkage p coincides with its corresponding lower joint. Figure 3 shows the bipolar Gaussian joint confidence map for the right knee-right ankle linkage.

To get a strong orientation and position representation of the body parts, PAF [25] are used to represent the body part linkages. The 2D vector field for the right knee-right ankle body linkage is shown in Fig. 3. The PAF ground truth of body part p of the individual, i, ${}^*L^p_i$, at any pixel $\boldsymbol{x}$ depends on whether or not the pixel is on the part link defined by a region along the vector from the high joint to the low joint is defined

Figure 3. Joint confidence and PAF for right knee-right ankle body linkage

with length l_p and width σ_p in pixels. The PAF vector field ${}^{*}L_i^p$ is computed as follows:

$$ {}^{*}L_i^p = \begin{cases} \boldsymbol{v} = \dfrac{\boldsymbol{x}_{jl,i}^p - \boldsymbol{x}_{jh,i}^p}{\left\| \boldsymbol{x}_{jl,i}^p - \boldsymbol{x}_{jh,i}^p \right\|_2} & \left\{ \begin{array}{l} 0 \le \boldsymbol{v} \cdot (\boldsymbol{x} - \boldsymbol{x}_{jh,i}^p) \le l_p, \text{ and} \\ 0 \le \left| \boldsymbol{v}_{\perp} \cdot (\boldsymbol{x} - \boldsymbol{x}_{jh,i}^p) \right| \le \sigma_p \end{array} \right\} \\ 0 & \text{otherwise} \end{cases} \quad (5) $$

where, $\boldsymbol{v}_{\perp}$ is the unit 2D vector perpendicular to $\boldsymbol{v}$. If more than one person is visible in an image, then the PAF ground truth of any body-part p in the image is the average of all the PAFs from the visible persons

$$ {}^{*}L^p = \frac{1}{n_p} \sum_i {}^{*}L_i^p \quad (6) $$

where n_p denotes the number of body parts detected in the input image.

System Training:

The proposed model was implemented using the Keras framework [27] with TensorFlow [28] as the backend engine. Multiple stochastic gradient descent optimizers are used to optimize the neural network with a total; epoch of 43. The learning rate for the network training was set to 2×10^{-4}. The system training performance is shown in Fig. 4 in the form of decay in system loss (as defined by Eq. (2)) with training epochs. The model was trained with 52,597 image samples from the COCO dataset on an Intel Xeon™ (6-Cores, 1.8 GHz) workstation with 32GB RAM and NVIDIA GTX1080 GPU. The training process lasted 129 hours and delivered a loss of joint confidence map of 37.72 in 43 epochs.

HUMAN POSE ESTIMATION

After the confidence map of body parts has been generated by the trained network, the human body pose estimation process takes charge of associating body parts to their respective persons. The pose estimation process takes place in two consecutive steps, (1) Body Parts Parsing: assemble part link with its associated joints, and (2) Individual Parsing: assembling all body parts to individual body skeleton.

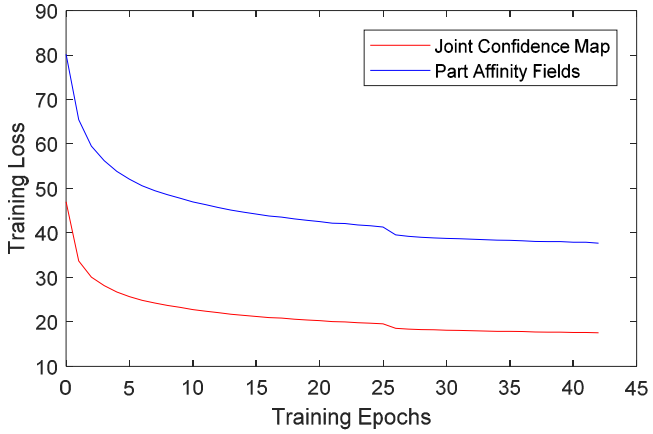

Figure 4. Model training performance

Body Parts Parsing:

In this process, the association of the high joints, low joints, and body part links are determined by computing the integral of the PAF between the two joint candidates. A voting mechanism has been used to find the best joint-link pairs. For the high joint and low joint pair candidate positions, $\boldsymbol{x}_{jh,m}^p, \boldsymbol{x}_{jl,n}^p$, the associated voting score can be calculated as follows:

$$ V_{m,n}^p = \int_{u=0}^{u=1} L^p\left(\boldsymbol{x}(u)\right) \cdot \frac{\boldsymbol{x}_{jh,n}^p - \boldsymbol{x}_{jl,m}^p}{\left\| \boldsymbol{x}_{jh,n}^p - \boldsymbol{x}_{jl,m}^p \right\|} du \quad (7) $$

$$ \boldsymbol{x}(u) = (1-u) \cdot \boldsymbol{x}_{jh,n}^p + u \cdot \boldsymbol{x}_{jl,m}^p $$

where $\boldsymbol{x}(u)$ denotes interpolated position between the high joint and low joint. The input to the p^{th} body part parsing is a set of the high joints peaks sets X_{jh}^p, low joints peaks sets X_{jl}^p and PAF sets L^p, which can be expressed as $\{X_{jh}^p, X_{jl}^p, L^p\}$. The output of the body part parsing is a set of high joint and low joint pairs, $\{(\boldsymbol{x}_{jh,n}^p, \boldsymbol{x}_{jl,m}^p) : n \in \{1,...,N_{jh}^P\}, m \in \{1,...,M_{jl}^P\}\}$. For each part p, a variable $C_{m,n}^p \in \{0,1\}$ is used to indicate whether the high joint and low joint pairs $\left(\boldsymbol{x}_{jh,m}^p, \boldsymbol{x}_{jl,n}^p\right)$ are connected or

Figure 5. Body part parsing: (A) Estimated high joint location, (B) Estimated low joint location, (C) Both Joints after body part parsing, (D) Skeleton output after individual parsing

not. The body parsing process aims to maximize the total voting score for connecting joint pairs via the following:

$$\begin{aligned} \max_{C^p_{m,n}} V_C^P &= \max_{C^p_{m,n}} \sum_m \sum_n V^p_{m,n} \cdot C^p_{m,n} \\ \forall m \in \{1,...,M^P_{jl}\}, &\sum_n C^p_{m,n} \le 1 \\ \forall n \in \{1,...,N^P_{jh}\}, &\sum_m C^p_{m,n} \le 1 \end{aligned} \tag{8}$$

The equation ensures that no joints are used to construct more than one body part.

Individual Parsing:

After the body part parsing, the individual parsing assembles all the body parts to form individual skeletons by assembling rules $A^{p,q}_{l,h} \in \{0,1\}$. These rules indicate whether the low joint of p can be connected to the high joint of q and then merged as one joint. Another variable store whether the low joints of p, $\boldsymbol{x}^p_{jl}$ are merged with the high joints of q, $\boldsymbol{x}^q_{jh}$. A voting score $V^{p,q}_{l,h}$ is then used to determine the possibility of the two joints merging into one joint,

$$V^{p,q}_{l,h} = -\frac{J^p\left(\boldsymbol{x}^p_{jl,l}\right) \cdot J^q\left(\boldsymbol{x}^q_{jh,h}\right)}{\left\| \boldsymbol{x}^p_{jl,l} - \boldsymbol{x}^q_{jh,h} \right\|_2} \tag{9}$$

Here, the negative sign is assigned to generate positive scores due to the negative values of the predicted confidence map of the low joints. The score will decrease as the joints grow further away. The only exception to this rule applies to the neck joints where five different body linkages get connected with their high joints,

$$'V^{p,q}_{\mathrm{h,h}} = \frac{J^p\left(\boldsymbol{x}^p_{jh,h}\right) \cdot J^q\left(\boldsymbol{x}^q_{jh,h}\right)}{\left\| \boldsymbol{x}^p_{jh,h} - \boldsymbol{x}^q_{jh,h} \right\|_2} \tag{10}$$

The multiple-person pose parsing process becomes one optimization problem to maximize the total assembling score,

$$\begin{aligned} \max_{C^{p,q}_{l,h}} V_C^{P,Q} &= \max_{C^{p,q}_{l,h}} \sum_l \sum_h (V^{p,q}_{l,h} \cdot A^{p,q}_{l,h} \cdot C^{p,q}_{l,h} + {}'V^{p,q}_{h,h} \cdot A^{p,q}_{h,h} \cdot C^{p,q}_{h,h}) \\ \forall A^{p,q}_{l,h} = 1, \forall l \in \{1,...,M^P_{jl}\}, &\sum_h C^{p,q}_{l,h} \le 1, \\ \forall A^{p,q}_{l,h} = 1, \forall h \in \{1,...,N^Q_{jh}\}, &\sum_l C^{p,q}_{l,h} \le 1 \end{aligned} \tag{11}$$

Here, $V_C^{P,Q}$ presents the possibility that joints could be merged into one single joint. The connected joint position is then refined by weighting the predicted position of the joints from the two body parts. The new connected joint position can be refined using the following expression:

$$\boldsymbol{x}^p_{jl,l} = \boldsymbol{x}^q_{jh,h} = \frac{\left|J^p\left(\boldsymbol{x}^p_{jl,l}\right)\right| \cdot \boldsymbol{x}^p_{jl,l} + \left|J^q\left(\boldsymbol{x}^q_{jh,h}\right)\right| \cdot \boldsymbol{x}^q_{jh,h}}{\left|J^p\left(\boldsymbol{x}^p_{jl,l}\right)\right| + \left|J^q\left(\boldsymbol{x}^q_{jh,h}\right)\right|}. \tag{12}$$

Once common joints are located, the connected links can be identified as 2D human pose skeletons.

EXPERIMENTS AND RESULTS

Results on COCO database

The COCO human keypoint dataset contains 57k images, out of which approximately 52k samples were used for the training dataset and the remaining 5k samples were used for validation. During the training of the neural network, the testing that was done on the validation set is performed at the end of each epoch. Figure 6 shows some samples from the validation set processed on the fully trained network. A good match in the predicted and the estimated pose was observed. As reported in the training section, a loss of 37.72 was obtained in the joint confidence map generation.

Runtime Analysis

The runtime of the proposed algorithm comprises of two major components: (1) the body parts detection process time, which is invariant to the number of persons shown in the image, with runtime complexity of *O(1)*; (2) the body parts

Figure 6. Pose Estimation results on COCO 2017 human keypoint validation dataset

assembling process time, whose runtime complexity is $O(n^2)$. Hence. the runtime increases with number of persons (n) in the input image. Compared to the body parts assembling process time, the body parts detection process time influences the total processing time even more. With a lesser number of sequential stages compared to [25], the proposed networks better take advantage of parallel computing and estimates the human pose in a more efficient way. By using a single GTX1080, the CNN takes 103.5 ms , compared to the CNN having taken 99.6 ms in [25]. The time of CNN computation decreases to 76.8 ms by applying the network on two GTX1080s. The body parsing takes 0.61 ms and does not change with the number of GPUs deployed.

CONCLUSION AND FUTURE WORK

This paper presented one parallel, multi-branch deep neural network architecture along with a corresponding post-processing method for multiple person 2D-pose estimation in the monocular image. Each trained branch (CNN ensemble) of the proposed neural network was trained to detect one specific body part (associated joints and link) for the human skeleton model and delivered an overall detection loss of 37.72. The highly parallel architecture achieves similar performance compared to the previous deep neural network architecture [25] but runs faster on a parallel computing devices such as GPU. This saves time for post-processing and the ensuing higher level tasks. This neural network is also highly adaptable for tasks aimed at specific body parts since branches of the last stage can be deployed independently. This feature also helps save storage and computing resources without sacrificing performance.

The proposed system will be augmented with multispectral imagery to enable the detection of human poses in various lighting conditions. Efforts are being made to improve the network architecture and post-processing algorithms to achieve better efficiency and faster runtime for purposes of deployment on embedded systems. The proposed system will be deployed on an autonomous mobile robotic platform to assist with the search and rescue of casualties in disaster management and war-like scenarios.

ACKNOWLEDGMENT

This work is supported in part by the US Army Medical Research & Materiel Command's Telemedicine & Advanced Technology Research Center (TATRC), under Contract No. W81XWH-16-C-0062. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

REFERENCES

[1] Gong, W., Zhang, X., Gonzàlez, J., Sobral, A., Bouwmans, T., Tu, C., and Zahzah, E., 2016, "Human Pose Estimation from Monocular Images: A Comprehensive Survey," Sensors (Basel)., **16**(22), pp. 1–39.

[2] Toshev, A., "DeepPose: Human Pose Estimation via Deep Neural Networks."

[3] Tompson, J., Jain, A., LeCun, Y., and Bregler, C., 2014, "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation," Adv. Neural Inf. Process. Syst., pp. 1799--1807.

[4] Li, S., Liu, Z.-Q., and Chan, A. B., 2014, "Heterogeneous Multi-Task Learning for Human Pose Estimation with Deep Convolutional Neural Network."

[5] Agarwal, A., and Triggs, B., 2006, "A Local Basis Representation for Estimating Human Pose from Cluttered Images," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), **3851 LNCS**, pp. 50–59.

[6] Jang, C., and Jung, K., 2008, "Human Pose Estimation Using Active Shape Models," Proc. World Acad. Sci. …, **2**(8), pp. 312–316.

[7] Shotton, J., Girshick, R. B., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., and Blake, A., 2012, "Efficient Human Pose Estimation from Single Depth Images.," IEEE Trans. Pattern Anal. Mach. Intell., **35**(12), pp. 2821–2840.

[8] Ye, M., Wang, X., Yang, R., Ren, L., and Pollefeys, M., 2011, "Accurate 3D Pose Estimation from a Single Depth Image," IEEE Int. Conf. Comput. Vis., pp. 731–738.

[9] Iqbal, U., Milan, A., and Gall, J., 2016, "PoseTrack: Joint Multi-Person Pose Estimation and Tracking."

[10] Dalal, N., and Triggs, W., 2004, "Histograms of Oriented Gradients for Human Detection," 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR05, **1**(3), pp. 886–893.

[11] Andriluka, M., Roth, S., and Schiele, B., 2009, "Pictorial Structures Revisited People Detection and Articulated Pose Estimation," *Computer Vision and Pattern Recognition, 2009*, pp. 1014–1021.

[12] Gkioxari, G., Hariharan, B., Girshick, R., and Malik, J., 2014, "R-CNNs for Pose Estimation and Action Detection," pp. 1–8.

[13] Fragkiadaki, K., Levine, S., Felsen, P., and Malik, J., 2015, "Recurrent Network Models for Human Dynamics," 2015 IEEE Int. Conf. Comput. Vis., pp. 4346–4354.

[14] Luo, P., Wang, X., and Tang, X., 2013, "Pedestrian Parsing via Deep Decompositional Network," Proc. IEEE Int. Conf. Comput. Vis., pp. 2648–2655.

[15] Dantone, M., Gall, J., Leistner, C., and Van Gool, L., 2013, "Human Pose Estimation Using Body Parts Dependent Joint Regressors," 2013 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3041–3048.

[16] Ouyang, W., Chu, X., and Wang, X., 2014, "24. Multi-Source Deep Learning for Human Pose Estimation,"

2014 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2337–2344.

[17] Chen, X., and Yuille, A., 2014, "Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations," pp. 1–9.

[18] Wei, S. E., Ramakrishna, V., Kanade, T., and Sheikh, Y., 2016, "Convolutional Pose Machines," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732.

[19] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B., 2016, "Deepercut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), **9910 LNCS**, pp. 34–50.

[20] Iqbal, U., and Gall, J., 2016, "Multi-Person Pose Estimation with Local Joint-to-Person Associations," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), **9914 LNCS**, pp. 627–642.

[21] Fang, H. S., Xie, S., Tai, Y. W., and Lu, C., 2017, "RMPE: Regional Multi-Person Pose Estimation," Proc. IEEE Int. Conf. Comput. Vis., **2017–Octob**, pp. 2353–2362.

[22] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., and Murphy, K., 2017, "Towards Accurate Multi-Person Pose Estimation in the Wild," pp. 4903–4911.

[23] Ren, S., He, K., Girshick, R., and Sun, J., 2017, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., **39**(6), pp. 1137–1149.

[24] Jin, S., Ma, X., Han, Z., Wu, Y., Yang, W., Liu, W., Qian, C., and Ouyang, W., "Towards Multi-Person Pose Tracking : Bottom-up and Top-down Methods," (2), pp. 4–7.

[25] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y., 2016, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields."

[26] Simonyan, K., and Zisserman, A., 2014, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14.

[27] Chollet, F., and Others, 2015, "Keras."

[28] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X., 2016, "TensorFlow: A System for Large-Scale Machine Learning," *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, USENIX Association, p. 44.