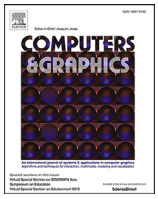

Special Section on SIBGRAPI 2019

Human pose regression by combining indirect part detection and contextual information

Diogo C. Luvizon[a,b,*], Hedi Tabia[a,c], David Picard[a,d]

[a] ETIS UMR 8051, ENSEA, CNRS, Paris Seine University, F-95000 Cergy, France
[b] Advanced Technologies, Samsung Research Institute, Campinas, Brazil
[c] IBISC, Univ. dÉvry Val-dÉssonne, Université Paris Saclay, France
[d] LIGM, UMR 8049, École des Ponts, UPE, Champs-sur-Marne, France

ARTICLE INFO

Article history:
Received 15 April 2019
Revised 4 September 2019
Accepted 7 September 2019
Available online 11 September 2019

Keywords:
Human pose estimation
Neural nets
Computer vision

ABSTRACT

In this paper, we tackle the problem of human pose estimation from still images, which is a very active topic, specially due to its several applications, from image annotation to human-machine interface. We use the *soft-argmax* function to convert feature maps directly to body joint coordinates, resulting in a fully differentiable framework. Our method is able to learn heat maps representations indirectly, without additional steps of artificial ground truth generation. Consequently, contextual information can be included to the pose predictions in a seamless way. We evaluated our method on two challenging datasets, the Leeds Sports Poses (LSP) and the MPII Human Pose datasets, reaching the best performance among all the existing regression methods. Source code available at: https://github.com/dluvizon/pose-regression.

1. Introduction

Human pose estimation from still images is a hard task since the human body is strongly articulated, some parts may not be visible due to occlusions or low quality images, and the visual appearance of body parts can change significantly from one pose to another. Classical methods use keypoint detectors to extract local information, which are combined to build pictorial structures [1]. To handle difficult cases of occlusion or partial visualization, contextual information is usually needed to provide visual cues that can be extracted from a broad region around the part location [2] or by interaction among detected parts [3]. In general, pose estimation can be seen from two different perspectives, namely as a correlated part detection problem or as a regression problem. Detection based approaches commonly try to detect keypoints individually, which are aggregated in post-processing stages to form one pose prediction. In contrast, methods based on regression use a function to map directly input images to body joint positions.

In the last few years, pose estimation have gained attention with the breakthrough of deep Convolutional Neural Networks (CNN) [4] alongside consistent computational power increase. This can be seen as the shift from classical approaches [5,6] to deep architectures. In many recent works from different domains, CNN based methods have overcome classical approaches by a large margin [7]. A key benefit from CNN is that the full pipeline is differentiable, allowing end-to-end learning. In the context of human pose estimation, the first methods using deep neural networks tried to do regression directly by learning a non-linear mapping function from RGB images to joint coordinates [4]. By contrast, the majority of the methods in the state of the art tackle pose estimation as a detection problem by predicting heat maps that correspond to joint locations [8,9], or even by exploiting additional tasks such as semantic body segmentation [10]. In such methods, the ground truth is artificially generated from joint positions, generally as a 2D Gaussian distribution centered on the joint location, while the context information is implicitly learned by the hidden convolutional layers.

Despite achieving state-of-the-art accuracy on 2D pose estimation, detection based approaches have some limitations. For example, such methods rely on additional steps to convert heat maps to joint positions, usually by applying the *argmax* function, which is not differentiable, breaking the learning chain on neural networks. Additionally, the precision of predicted keypoints is proportional to that of the heat maps resolution, which leads the top ranked methods [8,11] to high memory consumption and high computational requirements.

On the other hand, regression based methods are conceptually more adapted to 2D and 3D scenarios and can be used indistinctly

* Corresponding author at: ETIS UMR 8051, ENSEA, CNRS, Paris Seine University, F-95000 Cergy, France.
E-mail address: diogo.luvizon@ensea.fr (D.C. Luvizon).

https://doi.org/10.1016/j.cag.2019.09.002
0097-8493/

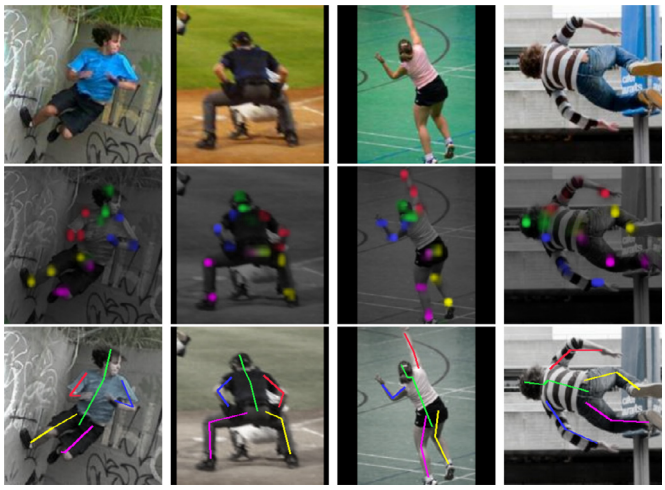

Fig. 1. Test samples from the Leeds Sports Poses (LSP) dataset. Input image (top), the predicted part-based maps encoded as RGB image for visualizasion (middle), and the regressed pose (bottom). Corresponding human limbs have the same colors in all images. This figure is better seen in color.

on both cases [12]. However, the regression function map is sub-optimally learned, resulting in lower scores when compared with detection based approaches. In this paper, we aim at solving this problem by bridging the gap between detection and regression based methods. We propose to replace the *argmax* function, used to convert heat maps into joint locations, by the *soft-argmax* function, which keeps the properties of specialized part detectors while being fully differentiable. The idea of *soft-argmax* was previously introduced by Finn et al. [13] in order to convert the highest response from a feature map to its coordinates. Differently from our work, in [13] the output of *soft-argmax* is not explicitly supervised. More recently, the *soft-argmax* was also used to guide local features extraction [14] and to perform 3D human pose estimation in [15], which is a parallel work to ours. With our solution based on *soft-argmax*, we are able to explore contextual information while optimizing our network from end-to-end using regression losses, *i.e.*, from input RGB images to final (x, y) body joint coordinates.

The contributions of our work are the following: *first*, we present a human pose regression approach from still images based on the *soft-argmax* function, resulting in an end-to-end trainable method which does not require artificial heat maps generation for training. *Second*, the proposed method can be trained using an insightful regression loss function, which is directly linked to the error distance between predicted and ground truth joint positions. *Third*, in the proposed architecture, contextual information is directly accessible and is easily aggregated to the final predictions. *Finally*, the accuracy reached by our method surpasses that of regression methods and is close to that of state-of-the-art detection methods, despite using a much smaller network. Some examples of our regressed poses are shown in Fig. 1.

The rest of this paper is divided as follows. In the next section, we present a review of the most relevant related work. The proposed method is presented in Section 3. In Section 4, we show the experimental evaluations, followed by our conclusions in Section 5.

2. Related work

Several approaches for human pose estimation have been presented for both 2D [16] and 3D [17,18] scenarios, as well as for video sequences [19–21]. Among classical methods, Pictorial Structures [22] and poselet-based features [23] have been widely used in the past. In this section, due to the limited space, we focus on CNN based methods that are more related to our work i.e., 2D human pose estimation from single frames. We briefly refer to the most recent works, splitting them as regression based and detection based approaches.

Regression based approaches. Some methods tackle pose estimation as a keypoint regression problem. One of the first regression approaches was proposed by Toshev and Szegedy [4] as a holistic solution based on cascade regression for body part detection, where individual joint positions are recursively improved, taking the full frame as input. Pfister et al. [24] proposed the Temporal Pose ConvNet to track upper body parts, and Carreira et al. [25] proposed the Iterative Error Feedback by injecting the prediction error back to the input space, improving estimations recursively. The handle the difficult cases of complex human poses, Rogez et al. [26] proposed the LCR network, on which each person is first *localized*, then *classified* according to a set of anchor poses, and finally the pose is *regressed*. The drawback of this method is the elevated number of pose anchors required to achieve reliable results. Recently, Sun et al. [12] proposed a structured bone based representation for human pose, which is statistically less variant than absolute joint positions and can be indistinctly used for both 2D and 3D representations. However, the method requires converting pose data to the relative bone based format. Moreover, those results are all outperformed by detection based methods.

Detection based approaches. Pischulin et al. [27] proposed DeepCut, a graph cutting algorithm that relies on body parts detected by DeepPose [4]. This method has been improved in [28] by replacing the previous CNN by a deep Residual Network (ResNet) [29], resulting in very competitive accuracy results, specially on multi-person detection. Semantic part based detection [30] is another possibility for human pose estimation, but it requires additional data annotation.

Several methods have shown significant improvements on accuracy by using fully convolutional models to generate belief maps (or heat maps) for joint probabilities [8,9,11,31,32]. For example, Bulat and Tzimiropoulos [9] proposed a two-stages CNN for coarse and fine heat map regression using pre-trained models, and following the tendency of deeper models with residual connections, Newell et al. [8] proposed a stacked hourglass network with convolutions in multi-level features, allowing reevaluation of previous estimations due to a stacked block architecture with many intermediate supervisions. The part-based learning process can benefit from intermediate supervision because it acts as constraints on the lower level layers. As a result, the feature maps on higher levels tend to be cleaner. More recently, the stacked hourglass network have been extended to more complex variations. For example, Chu et al. [11] proposed a Conditional Random Field (CRF) based on attention maps, and Yang et al. [33] studied variations of internal pyramids in multiple levels of each hourglass. To cope with unrealistic predictions, adversarial network have been used [34,35]. Despite their elevated memory consumption, these methods provide to our knowledge state-of-the-art performance.

All the previous methods that are based on detection need additional steps on training to produce artificial ground truth from joint positions, which represent an additional processing stage and additional hype-parameters, since the ground truth heat maps have to be defined by hand. On evaluation, the inverse operation is required, *i.e.*, heat maps have to be converted to joint positions, generally using the *argmax* function. Consequently, in order to achieve good precision, predicted heat maps need reasonable spacial resolution, as proposed in [8], which can translate into an elevated computational cost and memory usage. In order to provide an alternative to heat maps based approaches, we present our framework in the following section.

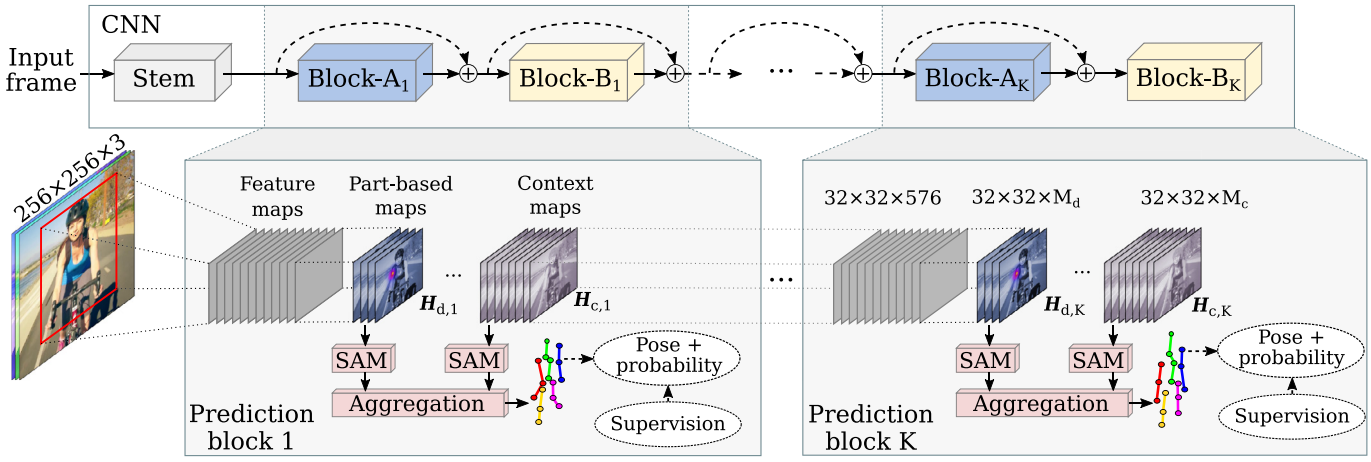

Fig. 2. Overview of the proposed approach for pose regression.

3. Proposed method

The proposed approach is an end-to-end trainable network which takes as input RGB images and outputs two vectors: the probability $\mathbf{p}_n$ of joint n being in the image and the regressed joint coordinates $\mathbf{y}_n = (x_n, y_n)$, where $n = \{1, 2, \ldots, N_J\}$ is the index of each joint and N_J is the number of joints. In what follows, we first present the global architecture of our method, and then detail its most important parts.

3.1. Network architecture

An overview of the proposed method is presented in Fig. 2. Our approach is based on a convolutional neural network essentially composed of three parts: one entry flow, block-A and block-B. The role of the stem is to provide basic feature extraction, while block-A and block-B provide refined features and body-part activation maps. One sequence of block-A and block-B is used to build one *prediction block*, which output is used as intermediate supervision during training. The full network is composed by the stem and a sequence of K prediction blocks. The final prediction is the output of the Kth prediction block. To predict the pose at each prediction block, we aggregate the 2D coordinates generated by applying *soft-argmax* to the part-based and contextual maps that are output by block-B. Similarly to recent approaches [8,11], on each prediction block we produce one estimation that is used as intermediate supervision, providing better accuracy and more stability to the learning process.

The proposed CNN model is partially based on Inception-v4 [36]. For block-A, we use a similar architecture as the Stacked Hourglass [8] replacing all the residual blocks by a residual separable convolution. Additionally, our approach increased the results from [8] with only three feature map resolutions, from 32×32 to 8×8, instead of the original five resolutions, from 64×64 to 4×4. This is possible because the *soft-argmax* is not directly dependent on the resolution of heat maps, since it performs a continuous regression, which is evidenced by our better results using lower resolution feature maps.

At each prediction stage, block-B is used to transform input feature maps into M_d *part-based detection maps* ($\mathbf{H}_d$) and M_c *context maps* ($\mathbf{H}_c$), resulting in $M = M_d + M_c$ heat maps. M_d corresponds to the number of joints N_J, and $M_c = N_c N_J$, where N_c is the number of context maps for each joint. The produced heat maps are projected back to the feature space and reintroduced to the network flow by a 1×1 convolution. Similar techniques have been used by many previous works [8,9,11], resulting in significant gain of performance. From the generated heat maps, our method predicts the joint locations and joint probabilities in the regression block, which has no trainable parameters. Details of block-B and the regression stage are shown in Fig. 3.

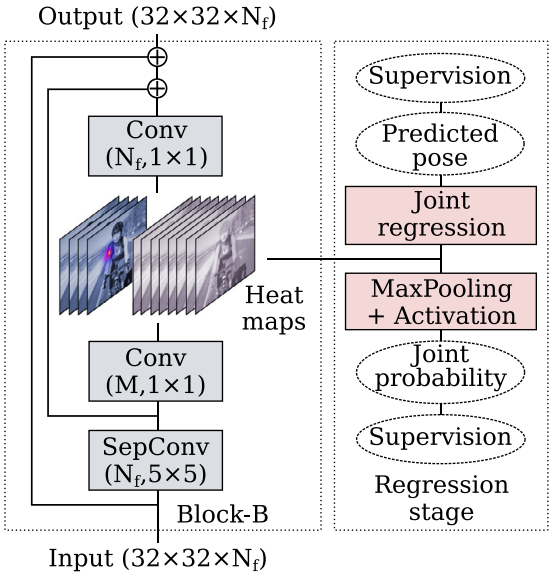

Fig. 3. Network architecture of block-B and an overview of the regression stage. The input is projected into M heat maps ($M_d + M_c$) which are then used for pose regression.

3.2. Proposed regression method

As presented in Section 2, traditional regression based methods use fully connected layers on feature maps and learn the regression mapping. However, this approach usually gives sub-optimal solutions. While state-of-the-art methods are overwhelmingly based on part detection, approaches based on regression have the advantages of providing directly the pose prediction as joint coordinates without additional steps or post-processing. In order to provide an alternative to detection based methods, we propose an efficient and fully differentiable way to convert heat maps directly to *(x, y)* coordinates, which we call *soft-argmax*. Additionally, the *soft-argmax* operation can be implemented as a CNN layer, as detailed in the next section.

3.2.1. Soft-argmax layer

Let us define the *softmax* operation on a single heat map $\mathbf{h} \in \mathbb{R}^{W \times H}$ as:

$$\Phi(\mathbf{h}_{i,j}) = \frac{e^{\mathbf{h}_{i,j}}}{\sum_{k=1}^{W}\sum_{l=1}^{H} e^{\mathbf{h}_{k,l}}}, \tag{1}$$

where $\mathbf{h}_{ij}$ is the value of heat map $\mathbf{h}$ at location *(i, j)*, and $W \times H$ is the heat map size. Contrary to the more common cross-channel softmax, we use here a spatial softmax that ensures each heat maps is normalized. Then, we define the *soft-argmax* as follows:

$$\Psi_d(\mathbf{h}) = \sum_{i=1}^{W}\sum_{j=1}^{H} \mathbf{W}_{i,j,d}\Phi(\mathbf{h}_{i,j}), \tag{2}$$

where d is a given component x or y, and $\mathbf{W}$ is a $W \times H \times 2$ weight matrix corresponding to the coordinates *(x, y)*. The matrix $\mathbf{W}$ can be expressed by its components $\mathbf{W}_x$ and $\mathbf{W}_y$, which are 2D discrete normalized ramps, defined as follows:

$$\mathbf{W}_{i,j,x} = \frac{i}{W}, \mathbf{W}_{i,j,y} = \frac{j}{H}. \tag{3}$$

Finally, given a heat map $\mathbf{h}$, the regressed location of the predicted joint is given by

$$\mathbf{y} = (\Psi_x(\mathbf{h}), \Psi_y(\mathbf{h}))^T. \tag{4}$$

This *soft-argmax* operation can be seen as a weighted average of points distributed on an uniform grid, with the weights being equal to the corresponding heat map. In order to integrate the *soft-argmax* layer into a deep network, we need its derivative with respect to $\mathbf{h}$:

$$\frac{\partial \Psi_d(\mathbf{h}_{i,j})}{\partial \mathbf{h}_{i,j}} = \mathbf{W}_{i,j,d}\frac{e^{\mathbf{h}_{i,j}}(\sum_{k=1}^{W}\sum_{l=1}^{H} e^{\mathbf{h}_{k,l}} - e^{\mathbf{h}_{i,j}})}{(\sum_{k=1}^{W}\sum_{l=1}^{H} e^{\mathbf{h}_{k,l}})^2}. \tag{5}$$

The *soft-argmax* function can thus be integrated in a trainable framework by using back propagation and the chain rule on Eq. (5). Moreover, from Eq. (5), we can see that the gradient is exponentially increasing for higher values, resulting in very discriminative response at the joint position.

The implementation of *soft-argmax* can be easily done with recent frameworks, such as TensorFlow, just by concatenating a spatial softmax followed by one convolutional layer with 2 filters of size $W \times H$, with fixed parameters according to Eq. (3).

Unlike traditional *argmax*, *soft-argmax* provides sub-pixel accuracy, allowing good precision even with very low resolution. Moreover, the *soft-argmax* operation allows to learn very discriminative heat maps directly from the *(x, y)* joint coordinates without explicitly computing artificial ground truth. Samples of heat maps learned by our approach are shown in Fig. 4.

3.2.2. Joint probability

Additionally to the joint locations, we estimate the joint probability $\mathbf{p}_n$, which corresponds to the probability of the nth joint being present in the image. The estimated joint probability is given by the sigmoid activation on the global max-pooling from heat map $\mathbf{h}_n$. Despite giving an additional piece of information, the joint probability does not depends on additional parameters and is computationally negligible, compared to the cost of convolutional layers.

3.2.3. Detection and context aggregation

Even if the correlation between some joints can be learned in the hidden convolutional layers, the joint regression approach is designed to locate body parts individually, resulting in low flexibility to learn from the context. For example, the same filters that give high response to images of a clean head, also must react positively to a hat or a pair of sunglasses. In order to provide multi-source information to the final prediction, we include in our framework specialized part-based heat maps and context heat maps,

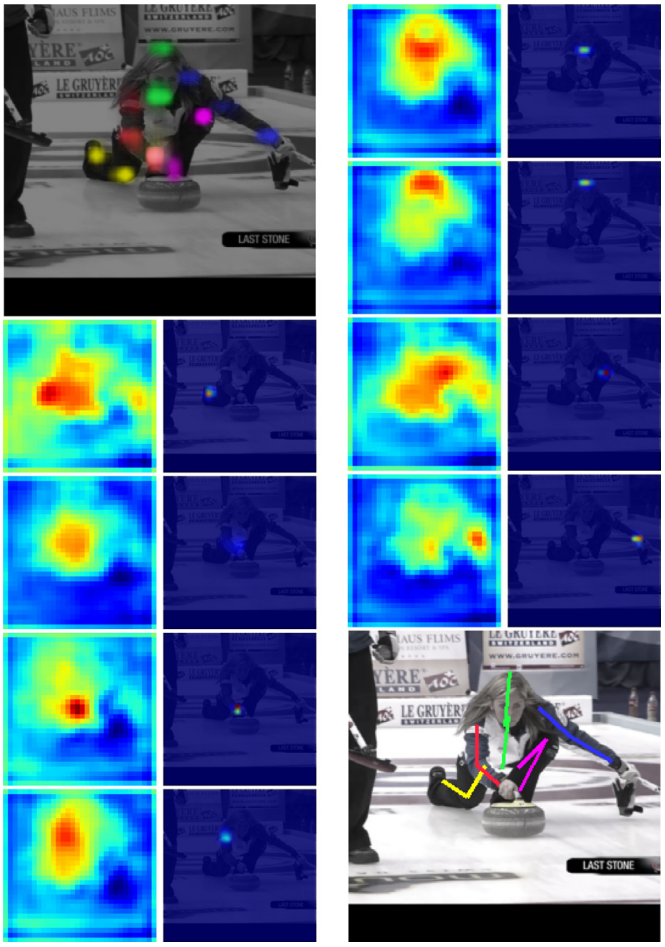

Fig. 4. Indirectly learned part-based heat maps from our method. All the joints encoded to RGB are shown in the first image (top-left corner) and the final pose is shown in the last image (bottom-right corner). On each column, the intermediate images correspond to the predicted heat maps before (left) and after (right) the Softmax normalization. The presented heat maps correspond to *right ankle*, *right hip*, *right wrist*, *right shoulder*, *upper neck*, *head top*, *left knee*, and *left wrist*.

which are defined as $\mathbf{H}_d = [\mathbf{h}^d_1, \ldots, \mathbf{h}^d_{N_j}]$ and $\mathbf{H}_c = [\mathbf{h}^c_{1,1}, \ldots, \mathbf{h}^c_{N_c,N_j}]$, respectively. Additionally, we define the joint probability related to each context map as $\mathbf{p}^c_{i,n}$, where $i = \{1, \ldots, N_c\}$ and $n = \{1, \ldots, N_j\}$.

Finally, the n^{th} joint position from detection and contextual information aggregated is given by:

$$\mathbf{y}_n = \alpha\mathbf{y}^d_n + (1-\alpha)\frac{\sum_{i=1}^{N_c}\mathbf{p}^c_{i,n}\mathbf{y}^c_{i,n}}{\sum_{i=1}^{N_c}\mathbf{p}^c_{i,n}}, \tag{6}$$

where $\mathbf{y}^d_n = \textit{soft-argmax}(\mathbf{h}^d_n)$ is the predicted location from the n^{th} part based heat map, $\mathbf{y}^c_{i,n} = \textit{soft-argmax}(\mathbf{h}^c_{i,n})$ and $\mathbf{p}^c_{i,n}$ are respectively the location and the probability for the i^{th} context heat map for joint n, and α is a hyper-parameter.

From Eq. (6) we can see that the final prediction is a combination of one specialized prediction and N_c contextual predictions pondered by their probabilities. The contextual weighted contribution brings flexibility, allowing specific filters to be more responsive to particular patterns. This aggregation scheme within the learning stage is only possible because we have the joint probability and position directly available inside the network in a differentiable way.

Table 1
Results considering different strategies for coordinates regression, evaluated using the PCKh@0.5 metric on the MPII *validation set*, single crop.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Simple *argmax*	95.8	91.3	86.7	82.4	85.8	75.5	76.7	85.3
soft-argmax w/o context	96.7	93.1	88.7	82.5	88.0	77.3	78.3	86.9
soft-argmax $\alpha = 0.9$	96.8	94.8	88.8	82.8	88.9	83.3	80.6	88.7
soft-argmax $\alpha = 0.8$	**96.8**	**95.2**	**89.0**	**82.9**	**89.2**	**84.6**	**80.9**	**89.1**

Table 2
Results on LSP test samples using the PCK measure at 0.2.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCK
Detection based methods								
Pishchulin et al. [5]	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1
Wei et al. [39]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat and Tzimiropoulos [9]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu et al. [11]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Yang et al. [33]	**98.3**	94.5	**92.2**	88.9	**94.4**	**95.0**	93.7	93.9
Chou et al. [35]	98.2	**94.9**	**92.2**	**89.5**	94.2	**95.0**	**94.1**	**94.0**
Regression based methods								
Carreira et al. [25]	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1
Our method	**97.5**	**93.3**	**87.6**	**84.6**	**92.8**	**92.0**	**90.0**	**91.1**

4. Experiments

We evaluate the proposed method on the very challenging MPII Human Pose [37] and Leeds Sports Poses (LSP) [38] datasets. The MPII dataset contains 25K images collected from YouTube videos, including around 28K annotated poses for training and 15K poses for testing. The annotated poses have 16 body joints, some of them are not present and others are occluded but can be predicted by the context. The LSP dataset is composed by 2000 annotated poses with up to 14 joint locations. The images were gethered from Flickr with sports people. The details about training the model and achieved accuracy results are given as follows.

4.1. Training

The proposed network was trained simultaneously on joints regression and joint probabilities. For joints regression, we use the elastic net loss function (L1 + L2):

$$L_{\mathbf{y}} = \frac{1}{N_J} \sum_{n=1}^{N_J} \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|_1 + \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|_2^2, \tag{7}$$

where $\mathbf{y}_n$ and $\hat{\mathbf{y}}_n$ are respectively the ground truth and the predicted n^{th} joint coordinates. In this case, we use directly the joint coordinates normalized to the interval [0,1], where the top-left image corner corresponds to (0,0), and the bottom-right image corner corresponds to (1,1).

For joint probability estimation, we use the binary cross entropy loss function on the joint probability $\mathbf{p}$:

$$L_{\mathbf{p}} = \frac{1}{N_J} \sum_{n=1}^{N_J} [(\mathbf{p}_n - 1)\ log\ (1 - \hat{\mathbf{p}}_n) - \mathbf{p}_n\ log\ \hat{\mathbf{p}}_n], \tag{8}$$

where $\mathbf{p}_n$ and $\hat{\mathbf{p}}_n$ are respectively the ground truth and the predicted joint probability.

We optimize the network using back propagation and the RMSProp optimizer, with batch size of 16 samples. For the MPII dataset, we train the network for 120 epochs. The learning rate begins at 10^{-3} and decreases by a factor of 0.4 when accuracy on validation plateaus. On the LSP dataset, we start from the model trained on MPII and fine-tuned it for more 70 epochs, beginning with learning rate $2 \cdot 10^{-5}$ and using the same decrease procedure. The full training of our network takes three days on the relatively outdated NVIDIA GPU Tesla K20 with 5GB of memory.

Data augmentation. We use standard data augmentation on both MPII and LSP datasets. Input RGB images are cropped and centered on the main subject with a squared bounding box, keeping the people scale (when provided), then resized to 256×256 pixels. We perform random rotations ($\pm 40^\circ$) and random rescaling from 0.7 to 1.3 to make the model more robust to image changes.

Parameters setup and ablation studies. Out network model is composed of eight prediction blocks ($K = 8$). We trained the network to regress 16 joints with 2 context maps for each joint ($N_j = 16,\ N_c = 2$). In the aggregation stage, we use $\alpha = 0.8$.

In order to provide insights about the chosen parameters, we performed some ablation studies as follows.

In Table 1, we evaluated the influence of the *soft-argmax* and the combination of contextual information on the precision of the method. The *soft-argmax* improves over a simple *argmax* by 1.6%, and the contextual maps improve precision by 2.2%. Not further significant improvement was noticed by using α lower than 0.8. The improvement is more relevant on more challenging joints, such as knees and ankles, which suggests that the contextual maps provide a complementary information to refine the specialized maps on difficult cases.

We also evaluated the execution time of our method comparing it with the stacked hourglass network [8], which is the most common baseline for detection approaches. Our method is able to perform predictions at 29.3 FPS (frames per second), while the stacked hourglass reached 18.3 FPS only, using the same framework and hardware (TensorFlow and NVIDIA GPU K20).

4.2. Results

LSP dataset. We evaluate our method on the LSP dataset using two metrics, the "Percentage of Correct Parts" (PCP) and the "Probability of Correct Keypoint" (PCK) measures. Our results compared to the state-of-the-art on the LSP dataset are present in Tables 2 and 3, respectively for PCK and PCP metrics. Our method achieves the best result among regression approaches. On the PCK measure, we outperform the results reported by Carreira et al. [25] (CVPR 2016) by 18.0%, which is the only regression method reported on this setup.

MPII dataset. On the MPII dataset, we evaluate our method using the "Single person" challenge [37]. The scores were computed by the providers of the dataset, since the test labels are not publicly available. As shown in Table 4, we reached a test score

Table 3
Results on LSP test samples using the PCP measure.

Method	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	PCP
Detection based methods							
Pishchulin et al. [5]	88.7	63.6	58.4	46.0	35.2	85.1	58.0
Wei et al. [39]	98.0	92.2	89.1	85.8	77.9	95.0	88.3
Bulat and Tzimiropoulos [9]	97.7	92.4	89.3	86.7	79.7	95.2	88.9
Chu et al. [11]	**98.4**	**95.0**	**92.8**	**88.5**	**81.2**	95.7	**90.9**
Regression based methods							
Carreira et al. [25]	95.3	81.8	73.3	66.7	51.0	84.4	72.5
Our method	**98.2**	**93.6**	**91.0**	**86.6**	**78.2**	**96.8**	**89.4**

Table 4
Comparison results with state-of-the-art methods on the MPII dataset on testing, using PCKh measure with threshold as 0.5 of the head segment length. Detection based methods are shown on top and regression based methods on bottom.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Detection based methods								
Pishchulin et al. [5]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Bulat and Tzimiropoulos [9]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al. [8]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Chu et al. [11]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al. [35]	98.2	96.8	92.2	88.8	91.3	89.1	84.9	91.8
Chen et al. [34]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang et al. [33]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Nie et al. [10][a]	**98.6**	**96.9**	**93.0**	**89.1**	**91.7**	**89.0**	**86.2**	**92.4**
Regression based methods								
Carreira et al. [25]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Sun et al. [12]	97.5	94.3	87.0	81.2	86.5	78.5	75.4	86.4
Our method	**98.1**	**96.6**	**92.0**	**87.5**	**90.6**	**88.0**	**82.7**	**91.2**

[a]Method using multi-task supervision with segmentation task (additional training).

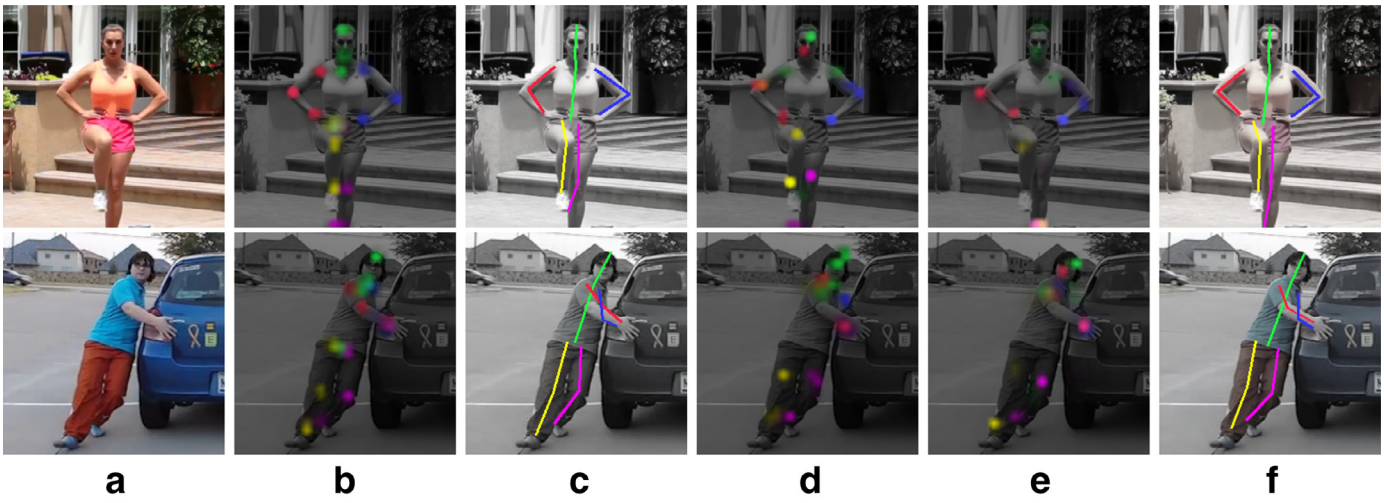

Fig. 5. Samples of context maps aggregated to refine predicted pose. Input image (a), part-based detection maps (b), predicted pose without context (c), two different context maps (d) and (e), and the final pose with aggregated predictions (f).

of 91.2%, which is 4.8% higher than the previous methods using regression.

Taking into account the competitiveness of the MPII Human Pose challenge, our score represents a very significant improvement over regression based approaches and a promising result compared to detection based methods. Moreover, our method is much simpler than the stacked hourglass network from Newell et al. [8] or its extensions [10,11,33–35]. For example, the size of the models [8,11,33] is 183 MB, 409 MB, and 217 MB, respectively, while our model requires only 58 MB. Due to limited memory resources, we were not able to re-train these models in our hardware. Despite that, we reach comparable results with a model that fits in much smaller GPUs.

4.3. Discussion

As suggested in Section 3.2.1, the proposed *soft-argmax* function acts as a constrain on the regression approach, driving the network to learn part-based detectors indirectly. This effect provides the flexibility of regression based methods, which can be easily integrated to provide 2D pose estimation to other applications such as 3D pose estimation or action recognition, while preserving the performance of detection based methods. Some examples of part-based maps indirectly learned by our method are shown in Fig. 4. As we can see, the responses are very well localized on the true location of the joints without explicitly requiring so.

The fact that the regressed coordinates of a given joint are influenced by all the pixels in the heat map could result in erroneous predictions in the case where multiple people are visible in the image. However, our method is trained with the target person centered in the cropped image, which makes our approach robust to the appearance of a second person in the corners (see an example in Fig. 4). In practice, a standard person detector [40] can be used to provide a well cropped bounding box around each person.

Additionally to the part-based maps, the contextual maps give extra information to refine the predicted pose. In some cases, the contextual maps provide strong responses to regions around the joint location. In such cases, the aggregation scheme is able to refine the predicted joint position. On the other hand, if the contextual map response is weak, the context reflects in very few changes on the pose. Some examples of predicted poses and visual contributions from contextual aggregation are shown in Fig. 5. The contextual maps are able to increase the precision of the predictions by providing complementary information, as we can see for the right elbows of the poses in Fig. 5.

5. Conclusion

In this work, we presented a new regression method for human pose estimation from still images. The method is based on the *soft-argmax* operation, a differentiable operation that can be integrated in a deep convolutional network to learn part-based detection maps indirectly, resulting in a significant improvement over the state-of-the-art scores from regression methods and very competitive results compared to detection based approaches. Additionally, we demonstrate that contextual information can be seamless integrated into our framework by using additional context maps and joint probabilities. As a future work, other methods could be build up to our approach to provide 3D pose estimation or human action recognition from pose in a fully differentiable way.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially founded by CNPq (Brazil) – Grant 233342/2014-1.

References

[1] Felzenszwalb PF, Huttenlocher DP. Pictorial structures for object recognition. Int J Comput Vis 2005;61(1):55–79.
[2] Fan X, Zheng K, Lin Y, Wang S. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2015.
[3] Yang Y, Baker S, Kannan A, Ramanan D. Recognizing proxemics in personal photos. In: Proceedings of the 2012 IEEE conference on computer vision and pattern recognition; 2012. p. 3522–9.
[4] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks. In: Proceedings of the computer vision and pattern recognition (CVPR); 2014. p. 1653–60.
[5] Pishchulin L, Andriluka M, Gehler PV, Schiele B. Strong appearance and expressive spatial models for human pose estimation. In: Proceedings of the international conference on computer vision (ICCV); 2013. p. 3487–94.
[6] Ladicky L, Torr PHS, Zisserman A. Human pose estimation using a joint pixel-wise and part-wise formulation. In: Proceedings of the computer vision and pattern recognition (CVPR); 2013.
[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR); 2016.
[8] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: Proceedings of the European conference on computer vision (ECCV); 2016. p. 483–99.
[9] Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression. In: Proceedings of the European conference on computer vision (ECCV); 2016. p. 717–32.
[10] Nie X, Feng J, Zuo Y, Yan S. Human pose estimation with parsing induced learner. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2018.
[11] Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X. Multi-context attention for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017.
[12] Sun X, Shang J, Liang S, Wei Y. Compositional human pose regression. In: Proceedings of the IEEE international conference on computer vision (ICCV); 2017.
[13] Finn C, Tan XY, Duan Y, Darrell T, Levine S, Abbeel P. Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. CoRR 2015;abs/1509.06113.
[14] Yi KM, Trulls E, Lepetit V, Fua P. Lift: Learned invariant feature transform. In: Proceedings of the European conference on computer vision. Springer; 2016. p. 467–83.
[15] Nibali A, He Z, Morgan S, Prendergast L. 3D human pose estimation with 2d marginal heatmaps. CoRR 2018. arXiv: 1806.01484.
[16] Dantone M, Gall J, Leistner C, Gool LV. Human pose estimation using body parts dependent joint regressors. In: Proceedings of the computer vision and pattern recognition (CVPR); 2013. p. 3041–8.
[17] Ionescu C, Li F, Sminchisescu C. Latent structured models for human pose estimation. In: Proceedings of the international conference on computer vision (ICCV); 2011. p. 2220–7.
[18] Luvizon DC, Picard D, Tabia H. 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2018.
[19] Pfister T, Charles J, Zisserman A. Flowing convnets for human pose estimation in videos. In: Proceedings of the international conference on computer vision (ICCV); 2015.
[20] Insafutdinov E, Andriluka M, Pishchulin L, Tang S, Levinkov E, Andres B, et al. Arttrack: Articulated multi-person tracking in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017.
[21] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV); 2018.
[22] Andriluka M, Roth S, Schiele B. Pictorial structures revisited: People detection and articulated pose estimation. In: Proceedings of the computer vision and pattern recognition (CVPR); 2009. p. 1014–21.
[23] Pishchulin L, Andriluka M, Gehler P, Schiele B. Poselet Conditioned Pictorial Structures. In: Proceedings of the computer vision and pattern recognition (CVPR); 2013. p. 588–95.
[24] Pfister T, Simonyan K, Charles J, Zisserman A. Deep convolutional neural networks for efficient pose estimation in gesture videos. In: Proceedings of the Asian conference on computer vision (ACCV); 2014.
[25] Carreira J, Agrawal P, Fragkiadaki K, Malik J. Human pose estimation with iterative error feedback. In: Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 4733–42.
[26] Rogez G, Weinzaepfel P, Schmid C. LCR-Net: Localization-Classification-Regression for Human Pose. In: Proceedings of the conference on computer vision and pattern recognition (CVPR); 2017. https://hal.inria.fr/hal-01505085.
[27] Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler P, et al. DeepCut: joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016.
[28] Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Proceedings of the European conference on computer vision (ECCV); 2016.
[29] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR); 2016.
[30] Liang X, Gong K, Shen X, Lin L. Look into person: joint body parsing & pose estimation network and a new benchmark. IEEE Trans Pattern Anal Mach Intell 2018;41(4):871–85.
[31] Belagiannis V, Rupprecht C, Carneiro G, Navab N. Robust optimization for deep regression. In: Proceedings of the International Conference on Computer Vision (ICCV); 2015. p. 2830–8.
[32] Cao Z, Simon T, Wei S, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 1302–10. doi:10.1109/CVPR.2017.143.
[33] Yang W, Li S, Ouyang W, Li H, Wang X. Learning feature pyramids for human pose estimation. In: Proceedings of the IEEE international conference on computer vision (ICCV); 2017.
[34] Chen Y, Shen C, Wei X-S, Liu L, Yang J. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: Proceedings of the IEEE international conference on computer vision (ICCV); 2017.
[35] Chou C, Chien J, Chen H. Self adversarial training for human pose estimation. CoRR 2017. abs/1707.02439.
[36] Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR 2016. abs/1602.07261.

[37] Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2014.
[38] Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British machine vision conference; 2010.
[39] Wei S-E, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016.
[40] Redmon J, Farhadi A. Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017.