Human pose regression by combining indirect part detection and contextual information
Diogo C Luvizon, Hedi Tabia, David Picard

To cite this version:
Diogo C Luvizon, Hedi Tabia, David Picard. Human pose regression by combining indirect part detection and contextual information. Computers and Graphics, 2019, 85, pp.15–22. 10.1016/j.cag.2019.09.002 . hal-02314445

HAL Id: hal-02314445
https://hal.science/hal-02314445v1
Submitted on 20 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License
Human pose regression by combining indirect part detection and contextual information

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

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

A R T I C L E I N F O

Article history:
Received September 3, 2019

Keywords: Human pose estimation, Neural nets, Computer vision.

A B S T R A C T

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.

© 2019 Elsevier B.V. All rights reserved.

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.

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

Anonymous C&G submission

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 app-

Fig. 1: Test samples from the Leeds Sports Poses (LSP) dataset. Input image (top), the predicted part-based maps encoded as RGB image for visualization (middle), and the regressed pose (bottom). Corresponding human limbs have the same colors in all images. This figure is better seen in color.
proaches 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 co-
dordinates [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 seg-
mentation [10]. In such methods, the ground truth is artificially
generated from joint positions, generally as a 2D Gaussian dis-
tribution centered on the joint location, while the context inform-
ation is implicitly learned by the hidden convolutional layers.

Despite achieving state-of-the-art accuracy on 2D pose esti-
ulation, detection based approaches have some limitations. For
eample, such methods rely on additional steps to convert heat
maps to joint locations, usually by applying the argmax func-
tion, which is not differentiable, breaking the learning chain on
neural networks. Additionally, the precision of predicted key-
points is proportional to that of the heat maps resolution, which
leads the top ranked methods [11, 8] to high memory consump-
tion and high computational requirements.

On the other hand, regression based methods are conceptu-
ally more adapted to 2D and 3D scenarios and can be used in-
distinctly 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 detec-
tion 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 spe-
cialized 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 out-
put of soft-argmax is not explicitly supervised. More recently,
the soft-argmax was also used to guide local features extrac-
tion [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 co-
dordinates.

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 di-
rectly linked to the error distance between predicted and ground
truth joint positions. Third, in the proposed architecture, con-
textual information is directly accessible and is easily aggre-
gated 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 sec-
tion, we present a review of the most relevant 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 pre-
vented for both 2D [16] and 3D [17, 18] scenarios, as well as for
video sequences [19, 20, 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 re-
gression based and detection based approaches.

Regression based approaches. Some methods tackle pose
estimation as a keypoint regression problem. One of the first re-
gression 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 im-
proved, taking the full frame as input. Pfister et al. [24] pro-
posed 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, im-
proving 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 re-
gressed. 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 indiscriminately 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 de-
tected 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 detec-
tion [30] is another possibility for human pose estimation, but it
requires additional data annotation.

Several methods have shown significant improvements on ac-
curacy by using fully convolutional models to generate belief
maps (or heat maps) for joint probabilities [11, 8, 31, 9, 32]. For
eexample, Bulat et al. [9] proposed a two-stages CNN for coarse
and fine heat map regression using pre-trained models, and fol-
2. Related work

Several approaches for human pose estimation have been pre-
vented for both 2D [16] and 3D [17, 18] scenarios, as well as for
video sequences [19, 20, 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 re-
gression based and detection based approaches.

Regression based approaches. Some methods tackle pose
estimation as a keypoint regression problem. One of the first re-
gression 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 im-
proved, taking the full frame as input. Pfister et al. [24] pro-
posed 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, im-
proving 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 re-
gressed. 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 indiscriminately 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 de-
tected 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 detec-
tion [30] is another possibility for human pose estimation, but it
requires additional data annotation.

Several methods have shown significant improvements on ac-
curacy by using fully convolutional models to generate belief
maps (or heat maps) for joint probabilities [11, 8, 31, 9, 32]. For
eexample, Bulat et al. [9] proposed a two-stages CNN for coarse
and fine heat map regression using pre-trained models, and fol-
lowing the tendency of deeper models with residual connec-
tions, 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 pro-
cess can benefit from intermediate supervision because it acts

Several approaches for human pose estimation have been pre-
vented for both 2D [16] and 3D [17, 18] scenarios, as well as for
video sequences [19, 20, 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 re-
gression based and detection based approaches.

Regression based approaches. Some methods tackle pose
estimation as a keypoint regression problem. One of the first re-
gression 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 im-
proved, taking the full frame as input. Pfister et al. [24] pro-
posed 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, im-
proving 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 re-
gressed. 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 indiscriminately 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 de-
tected 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 detec-
tion [30] is another possibility for human pose estimation, but it
requires additional data annotation.

Several methods have shown significant improvements on ac-
curacy by using fully convolutional models to generate belief
maps (or heat maps) for joint probabilities [11, 8, 31, 9, 32]. For
eexample, Bulat et al. [9] proposed a two-stages CNN for coarse
and fine heat map regression using pre-trained models, and fol-
lowing the tendency of deeper models with residual connec-
tions, 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 pro-
cess can benefit from intermediate supervision because it acts
as constraints on the lower layer levels. 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.

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 $p_n$ of joint $n$ being in the image and the regressed joint coordinates $y_n = (x_n, y_n)$, 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 $K$th 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 $32 \times 32$ to $8 \times 8$, instead of the original five resolutions, from $64 \times 64$ to $4 \times 4$. 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 ($H_d$) and $M_c$ context maps ($H_c$), 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 $1 \times 1$ convolution. Similar techniques have been used by many previous works [9, 8, 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.

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 suboptimal 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. Soft-argmax layer

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

\[ \Phi(h_{i,j}) = \frac{e^{h_{i,j}}}{\sum_k e^{h_{i,k}}}, \]

where \( h_{i,j} \) is the value of heat map \( h \) 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 map is normalized. Then, we define the soft-argmax as follows:

\[ \Psi_d(h) = \sum_l \sum_j W_{i,j,d} \Phi(h_{i,j}), \]

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

\[ W_{i,j,x} = \frac{i}{W}, \quad W_{i,j,y} = \frac{j}{H}. \]

Finally, given a heat map \( h \), the regressed location of the predicted joint is given by

\[ y = (\Psi_y(h), \Psi_x(h))^T. \]

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 \( h \):

\[ \frac{\partial \Psi_d(h_{i,j})}{\partial h_{i,j}} = W_{i,j,d} \frac{e^{h_{i,j}}}{\left( \sum_k e^{h_{i,k}} \right)^2}. \]

The soft-argmax function can thus be integrated in a trainable framework by using back propagation and the chain rule on equation (5). Moreover, from equation (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 equation (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 \( p_n \), which corresponds to the probability of the \( n^{th} \) joint being present in the image. The estimated joint probability is given by the sigmoid activation on the global max-pooling from heat map \( h_n \). Despite giving an additional piece of information, the joint probability does not depend 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, which are defined as \( H_d = \{ h_1, ..., h_N \} \) and \( H_c = \{ h_{11}, ..., h_{Nc} \} \), respectively. Additionally, we define the joint probability related to each context map as \( p_{i,n} \), where \( i = \{ 1, ..., N \} \) and \( n = \{ 1, ..., N_c \} \).

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

\[ y_n = \alpha y_n^d + (1 - \alpha) \frac{\sum_{i=1}^{N_c} p_{i,n} y_{i,n}}{\sum_{i=1}^{N_c} p_{i,n}}, \]

where \( y_n^d = \text{soft-argmax}(h_n^d) \) is the predicted location from the \( n^{th} \) part based heat map, \( y_{i,n}^c = \text{soft-argmax}(h_{i,n}^c) \) and \( p_{i,n} \) are respectively the location and the probability for the \( i^{th} \) context heat map for joint \( n \), and \( \alpha \) is a hyper-parameter.
From equation (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.

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 gathered 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):

$$L_y = \frac{1}{N_j} \sum_{n=1}^{N_j} ||y_n - \hat{y}_n||_1 + ||y_n - \tilde{y}_n||_2^2,$$

where $y_n$ and $\hat{y}_n$ 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 $p$:

$$L_p = \frac{1}{N_j} \sum_{n=1}^{N_j} [ (p_n - 1) \log (1 - \tilde{p}_n) - p_n \log \tilde{p}_n ],$$

where $p_n$ and $\tilde{p}_n$ 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 $256 \times 256$ 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. Our 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 $\alpha$ 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 PCP and PCK 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 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 [11, 35, 34, 33, 10]. For example, the size of the models [8], [11], and [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...
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 Tzimi. [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 |

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

| Method                          | Torso | Upper  | Lower  | Upper  | Fore- | 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 Tzimi. [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 Tzimi. [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]               | 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  |

* Method using multi-task supervision with segmentation task (additional training).
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 many 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.

![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.](image)

![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.](image)

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.

## 6. Acknowledgements

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

## References

[1] P. F. Felzenszwalb, D. P. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1) (2005) 55–79 (2005).

[2] X. Fan, K. Zheng, Y. Lin, S. Wang, Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 (June 2015).

[3] Y. Yang, S. Baker, A. Kannan, D. Ramanan, Recognizing proxemics in personal photos, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3522–3529 (June 2012).

[4] A. Toshev, C. Szegedy, DeepPose: Human Pose Estimation via Deep Neural Networks, in: Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653–1660 (2014).

[5] L. Pishchulin, M. Andriluka, P. V. Gehler, B. Schiele, Strong appearance and expressive spatial models for human pose estimation, in: International Conference on Computer Vision (ICCV), 2013, pp. 3487–3494 (2013).

[6] L. Ladicky, P. H. S. Torr, A. Zisserman, Human pose estimation using a joint pixel-wise and part-wise formulation, in: Computer Vision and Pattern Recognition (CVPR), 2013 (2013).

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (June 2016).

[8] A. Newell, K. Yang, J. Deng, Stacked Hourglass Networks for Human Pose Estimation, European Conference on Computer Vision (ECCV) (2016) 483–499 (2016).

[9] A. Bulat, G. Tzimiropoulos, Human pose estimation via Convolutional Part Heatmap Regression, in: European Conference on Computer Vision (ECCV), 2016, pp. 717–732 (2016).

[10] X. Nie, J. Feng, Y. Yao, S. Yan, Human pose estimation with parsing induced learner, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 (June 2018).
[11] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang. Multi-context attention for human pose estimation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 (July 2017).

[12] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, in: The IEEE International Conference on Computer Vision (ICCV), 2017 (Oct 2017).

[13] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, P. Abbeel, Learning visual feature spaces for robotic manipulation with deep spatial autoencoders., CoRR abs/1509.06113 (2015).

[14] K. M. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: Learned invariant feature transform, in: European Conference on Computer Vision, Springer, 2016, pp. 467–483 (2016).

[15] A. Nibali, Z. He, S. Morgan, L. Prendergast, 3d human pose estimation with 2d marginal heatmaps, CoRR abs/1806.01484 (2018). arXiv: 1806.01484.

[16] M. Dantone, J. Gall, C. Leistner, V. L. Gool, Human Pose Estimation Using Body Parts Dependent Joint Regressors, in: Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3041–3048 (June 2013).

[17] C. Ionescu, F. Li, C. Sminchisescu, Latent structured models for human pose estimation, in: International Conference on Computer Vision (ICCV), 2011, pp. 2220–2227 (Nov 2011).

[18] D. C. Luvizon, D. Picard, H. Tabia, 2d/3d pose estimation and action recognition using multitask deep learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 (June 2018).

[19] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: International Conference on Computer Vision (ICCV), 2015 (2015).

[20] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinek, B. Andres, B. Schiele, Arttrack: Articulated multi-person tracking in the wild, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 (July 2017).

[21] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: The European Conference on Computer Vision (ECCV), 2018 (September 2018).

[22] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, in: Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1014–1021 (June 2009).

[23] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Poselet Conditioned Pictorial Structures, in: Computer Vision and Pattern Recognition (CVPR), 2013, pp. 588–595 (June 2013).

[24] T. Pfister, K. Simonyan, J. Charles, A. Zisserman, Deep convolutional neural networks for efficient pose estimation in gesture videos, in: Asian Conference on Computer Vision (ACCV), 2014 (2014).

[25] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, Human pose estimation with iterative error feedback, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4733–4742 (June 2016).

[26] G. Rogez, P. Weinzaepfel, C. Schmid, LCR-Net: Localization-Regression-Classification for Human Pose, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2017 (2017), DOI: 10.1109/CVPR.2017.143.

[27] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele, DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (June 2016).

[28] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, DeepPerCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model, in: European Conference on Computer Vision (ECCV), 2016 (May 2016).

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (2016).

[30] X. Liang, K. Gong, X. Shen, L. Lin, Look into person: Joint body parsing & pose estimation network and a new benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence (2018).

[31] V. Belagiannis, C. Rupprecht, G. Carneiro, N. Navab, Robust optimization for deep regression, in: International Conference on Computer Vision (ICCV), 2015, pp. 2830–2838 (Dec 2015).

[32] Z. Cao, T. Simon, S. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1302–1310 (July 2017). doi:10.1109/CVPR.2017.143.

[33] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: The IEEE International Conference on Computer Vision (ICCV), 2017 (Oct 2017).

[34] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial posenet: A structure-aware convolutional network for human pose estimation, in: The IEEE International Conference on Computer Vision (ICCV), 2017 (Oct 2017).

[35] C. Chou, J. Chien, H. Chen, Self adversarial training for human pose estimation, CoRR abs/1707.02439 (2017).

[36] URL http://arxiv.org/abs/1707.02439.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-v4, inception-resnet and the impact of residual connections on learning, CoRR abs/1602.07261 (2016).

[38] URL http://arxiv.org/abs/1602.07261.

[39] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014 (June 2014).

[40] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in: Proceedings of the British Machine Vision Conference, 2010 (2010).

[41] J. Richardson, A. Farhadi, Yolo9000: Better, faster, stronger, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 (2016).
Conference on Computer Vision and Pattern Recognition (CVPR), 2017 (July 2017).