Journal of Ambient Intelligence and Humanized Computing (2021) 12:2339–2353
https://doi.org/10.1007/s12652-020-02347-7

ORIGINAL RESEARCH

Uniting holistic and part-based attitudes for accurate and robust deep human pose estimation

Faranak Shamsafar[1,2] · **Hossein Ebrahimnezhad**[1]

Received: 16 March 2020 / Accepted: 11 July 2020 / Published online: 28 July 2020

Abstract
Deep learning has been utilized in many intelligent systems, including computer vision techniques. Human pose estimation is one of the popular tasks in computer vision that has benefited from modern feature learning strategies. In this regard, recent advances propose part-based approaches since pose estimation based on parts can produce more accurate results than when the human shape is considered holistically as one unbreakable, but deformable object. However, in real-word scenarios, problems like occlusion and cluttered background make difficulties in part-based methods. In this paper, we propose to unite the two attitudes of the part-based and the holistic pose predictions to make more accurate and more robust estimations. These two schemes are modeled using convolutional neural networks as regression and classification tasks in order, and are combined in three frameworks: multitasking, series, and parallel. Each of these settings has its own advantages, and the experimental results on the LSP test set demonstrate that it is essential to observe subjects, both based on parts and holistically in order to achieve more accurate and more robust estimation of human pose in challenging scenarios.

Keywords Human pose estimation · Holistic prediction · Part-based prediction · Deep learning · Convolutional neural network

1 Introduction

In the computer vision community, human pose estimation is the art of automatically recovering the skeletal pose of a person from visual data. This can be very helpful in various industrial applications, such as surveillance, image retrieval, virtual or augmented reality, driving systems, etc. Real-world images introduce some main challenges for this goal, which are occlusion, cluttered background, high variability of human appearance, and high degree of articulation of a human pose.

In most of the recent work, human is considered based on the individual parts of the body and the visual data is processed in various regressor models to infer the pose (Belagiannis et al. 2015; Carreira et al. 2016; Li et al. 2015; Lifshitz et al. 2016; Sun et al. 2017; Zhou et al. 2016). In other words, the outputs of the model are real values, which represent body part locations or heatmaps, which show the probability of body parts existence. These approaches usually predict part locations with a suitable accuracy; however, they may get into trouble when the background is intensely cluttered or when a body part is severely occluded and not visible.

On the other hand, it is also possible to consider the human body as a whole (holistic) object, which can be deformed. This means that the parts of the human body are not taken into account individually. In this approach, however, the precise mapping from the space of RGB images to a high-DOF pose space is not straightforward. Namely, image description and handling the pose variance is highly critical in this method. This issue gains importance when there is no specific predefined limit in the image capturing process and in the pose configuration.

Therefore, both types of part-based and holistic predictions, have their own advantages and disadvantages, and they can complement each other. More precisely:

✉ Hossein Ebrahimnezhad
ebrahimnezhad@sut.ac.ir

Faranak Shamsafar
f_shamsafar@sut.ac.ir

[1] Computer Vision Research Laboratory, Electrical Engineering Faculty, Sahand University of Technology, Tabriz, Iran

[2] Present Address: WSI Institute for Computer Science, University of Tuebingen, Tuebingen, Germany

- The learning criteria in the holistic methods, unlike the part-based attitudes, is not directly a function of the displacements between individual body parts and their corresponding ground truth locations. Accordingly, the output of the holistic models would be a coarse estimation of the human pose, whereas the part-based frameworks try to approach to the precise individual joint locations.
- Part-based designs cannot predict a severely occluded joint since the visual information of the joint is lost. On the other hand, the holistic method is capable of predicting a probable location of an occluded part due to learning the overall holistic pose.
- When the background is highly cluttered or when there are similar objects in the background, false detection increase in the part-based frameworks. The holistic methods, on the contrary, do not focus on individual objects in the scene.
- In images with highly articulated poses, the part-based techniques may estimate a pose that is not anthropometrically valid for a human being. However, the problem of invalid pose prediction can be alleviated in the holistic methods as it is possible to define the whole human poses based on the real valid human poses.

Thus, by fusing the two attitudes of holistic and part-based estimations, we can obtain both more accurate and more robust pose prediction. Viewing from the perspective of our own experiences and observations, it is easy to notice that we do not recognize and observe human beings by only focusing on their individual body parts. In fact, our brain performs a type of attention to the whole body visual data of a person.

Based on these motivations, this work proposes the fusion of the two approaches of holistic and part-based, as classification and regression tasks in order, and this is our main contribution. The forms of fusion are threefold: multitasking, series, and parallel. An overview of our final method is depicted in Fig. 1, which uses multitasking and parallel fusion together (which are explained in Sect. 3). Unlike other part-based methods, the proposed uniting framework does not require feedback, cascading or repetition design, or other supplementary information, such as action label.

It should be noted that the input data to our system is one single frame. This type of data is the most challenging scheme in comparison to the multi-view or video sequence scenarios since it provides the least cues from the scene; no depth perception as in multi-view (Yan et al. 2020a) and no temporal information as in a video sequence is available. Yet, estimating the human pose based on one single frame is the most applicable case in many real-world applications, where there are specific device standards, particularly in the embedded systems, in terms of processing time, memory, and input resources. Enhancing pose estimation from still images can improve other cases as well (Yan et al. 2020b).

The rest of the paper is organized as follows: in Sect. 2, a summary of previous work on human pose estimation is presented. The proposed approaches are explained in detail in Sect. 3. In Sect. 4, the implementation details and experimental results are discussed. Finally, the paper concludes in Sect. 5.

2 Related work

There is an affluent research behind human pose estimation. Initially, the traditional approaches exploited handcrafted features like, HOG (Dalal and Triggs 2005), LBP (Ojala et al. 1994), etc. These features are fed into a model to search for a human pose by minimizing a matching function. The most prominent traditional models are

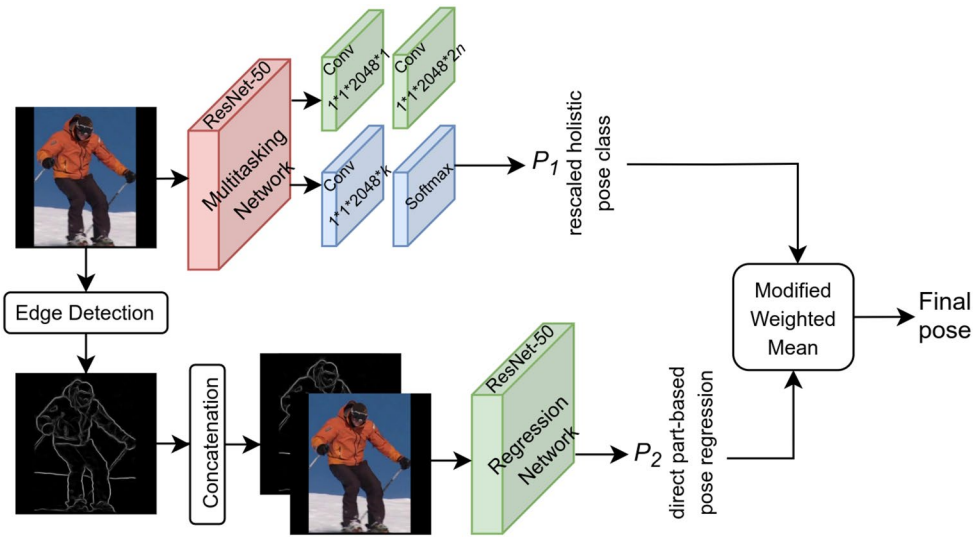

Fig. 1 An overview of our proposed approach: uniting the holistic and the part-based attitudes for human pose estimation

Deformable Part-based Models (DPMs) (Felzenszwalb et al. 2008, 2010b), which were suggested based on Pictorial Structures (Felzenszwalb and Huttenlocher 2005). These types of designs demonstrated state-of-the-art results prior to deep learning-based approaches.

While DPMs were formulated based on the individual body parts, other methods used a direct holistic mapping framework to estimate the human pose (Agarwal and Triggs 2006; Gavrila 2007; Mori and Malik 2002; Rogez et al. 2008; Shakhnarovich et al. 2003). However, these holistic perspectives were not considered for further research since, (1) they needed much stronger functions for mapping the image to the high variable human pose space, (2) they could only be used for heavily constrained conditions in laboratories, (3) there were a lack of data samples to get access to the desired variations, (4) there had to be a high resemblance between the training and the testing data in order to attain acceptable results, and (5) the method was not able to generalize to new unseen poses. Therefore, many researchers focused on part-based methods and proposed different variations of DPMs (Felzenszwalb et al. 2010a; Kokkinos 2012; Yang and Ramanan 2013).

After the rise of deep learning in computer vision applications, human pose estimation also took advantage of this modern promising technique. Some work like (Chen and Yuille 2014; Jain et al. 2014) only made use of learned features instead of handcrafted ones. Others, like (Tompson et al. 2014; Yang et al. 2016), trained the network in a way that it learns the pairwise relations between body parts as well. Recent methods take paradigms like feedbacking, cascading, repetition or larger receptive fields to achieve better recognition accuracies (Carreira et al. 2016; Wei et al. 2016). They all establish superiority with a large margin over the traditional approaches in human pose estimation in real-world images.

The aforementioned advances have one main idea in common, i.e. recovering human pose based on *individual body parts*. Actually, most of the design types have been put forward to overcome the problems of part-based attitude, which are occlusion, self-occlusion, false detection in cluttered backgrounds, double counting, and yielding anthropometrically invalid human poses.

Unlike these techniques, in Shamsafar and Ebrahimnezhad (2018), it is proposed to get a holistic prediction of human pose based on a classifier convolutional neural network. The method demonstrates competitive coarse pose estimation in comparison to the part-based schemes. The holistic pose prediction yields a coarse, but anthropometrically valid pose and it shows robustness in challenges like occlusion and cluttered backgrounds. The holistic pose estimation can be used in applications that demand a fast valid understating of the human pose, such as content-based image retrieval and scene understanding. Still, the accuracy of holistic prediction needs to be improved.

Note that human pose estimation can be coarsely estimated using other sources of information like radio tomographic imaging (RTI). Authors in Liu et al. (2014) have proposed to recognize fall detection using body pose. Depth information can also be utilized in coarse human pose estimation. For example, in Zavala-Mondragon et al. (2019), training a convolutional neural network is proposed to extract the actigraphy-related body pose information from depth images. The method is mainly toward clinical applications of estimating the human pose.

Based on what is discussed above, the two attitudes of holistic and part-based pose estimations have their own pros and cons. In this paper, we propose to use both of the attitudes to improve the performance of human pose estimation in monocular RGB images using convolutional neural networks.

3 Proposed approach

In this section, we first describe the holistic and the part-based pipelines, which are a classifier and a regressor convolutional neural network, respectively. Then, we propose to unite the two methods in three frameworks: multitasking, series, and parallel. Our final model, which is shown in Fig. 1, fuses the parallel and the multitasking frameworks. In the last Sect. 3.6 a strategy to assist the network for human pose estimation is suggested.

3.1 Holistic estimation

To this end, a classifier network is used similar to (Shamsafar and Ebrahimnezhad 2018). The fundamental concept in classification is the categorization of the input data into a predefined set of human pose classes. In Fig. 2, some pose classes are displayed, which are computed using the k-means++ algorithm on human pose data of the training

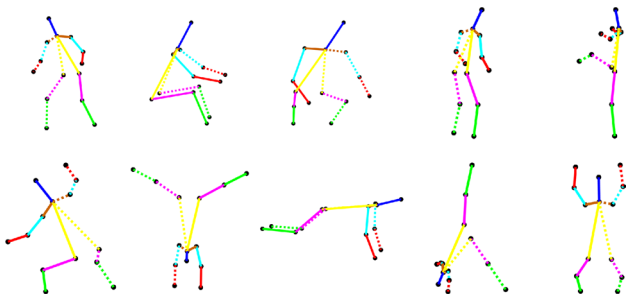

Fig. 2 Sample pose classes for pose classification in the holistic approach

image samples. The loss function of the classification (holistic) network is a multinomial logistic loss function. This method is considered as a holistic approach because the loss function penalizes the errors of misclassifying the pose classes, and not the individual joint displacements as in the part-based methods. Hence, any joint-based computations during training and testing is ignored and the human pose is considered as an entire shape that is capable of being deformed. For inference, the final pose is computed from the weighted mean of the t top pose classes. The weights are the class scores computed by the network. Additionally, a rescaling is required to fit the computed pose to the image. In our experiments, we have considered $t = 5$ and rescaling is performed using the height and the width of the input image. We refer the reader to Shamsafar and Ebrahimnezhad (2018) for more details.

3.2 Part-based estimation

While classifying human pose can be useful in providing a coarse and valid pose in challenging images, it cannot get closer to each individual joint location. Thus, in order to make the prediction of joint locations finer, a regression network is essential. After a network is modified in the last fully-connected layers to obtain a $1 \times 2n$ vector, the loss function for each sample is the mean of joint displacements using the Euclidean function:

$$L_{sample} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_i^{pr} - x_i^{gt})^2 + (y_i^{pr} - y_i^{gt})^2}, \quad (1)$$

where n is the number of joints, and (x_i^{pr}, y_i^{pr}) and (x_i^{gt}, y_i^{gt}) are the predicted and the ground truth locations for the i-th joint in order. In this method, unlike the described holistic estimation, all joint information play a direct role in minimizing the error function.

3.3 Multitasking fusion

A deep network can be trained for several tasks with various loss functions. This strategy is a replacement for training multiple individual networks. Here, we suggest using a network for both holistic and part-based pose predications. Specifically, the network performs both tasks of classification and regression. Two loss functions are defined for one network and other parameters are hard-shared during training. That means most of the layer weights are shared between these two tasks. Hard parameter sharing is advantageous in preventing the network from overfitting and it can be used in tasks that are semantically related. The architecture of the general hard parameter sharing and of our multitasking approach are illustrated in Figs. 3 and 4 in order.

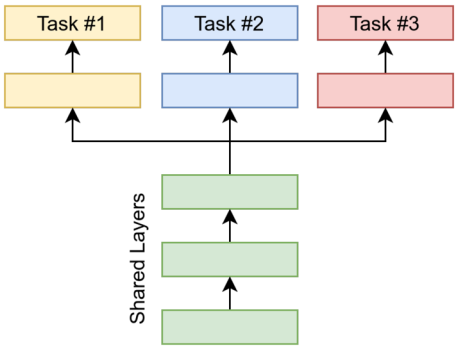

Fig. 3 Hard parameter sharing in multitask learning

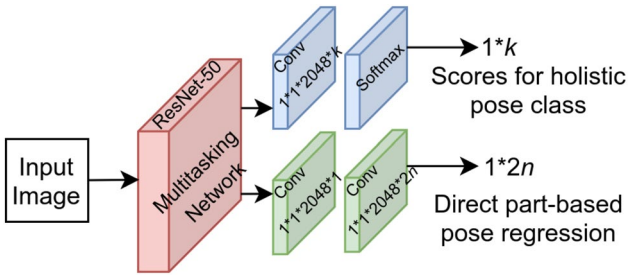

Fig. 4 The framework for the multitasking network. k and n indicate the number of pose classes and the number of joints in order

3.4 Series fusion

In the series technique, we suggest to initially estimate the probable locations of human body parts using the holistic method. Then, we use these predictions concatenated with the input image in the part-based network. The block diagram of the series method is displayed in Fig. 5. To this end, we first train a network as a human pose classifier. Then, the network is tested on unseen samples to compute the pose classes for them. One mask image is generated for each of the predicted joints as follows: After holistic pose computation, the mutual distances of all the predicted joints are computed. With a size as half of the maximum distance, a square mask centered on each joint location is built. Since we have defined the skeletal pose similar to the LSP dataset standard with 14 joints (Johnson and Everingham 2010), there is a total of 14 mask images, each showing a square-like neighborhood of the predicted joint. Next, these 14 joint masks with the 3 RGB channels of the input image are fed into a regressor network that directly predicts the 2D joint locations. As a result, the dimension of the input data to the regressor network would be $h \times w \times 17$, where h and w are the spatial dimensions of the image.

The motivation behind this fusion is to provide a type of attention about the neighborhood of the joints to the

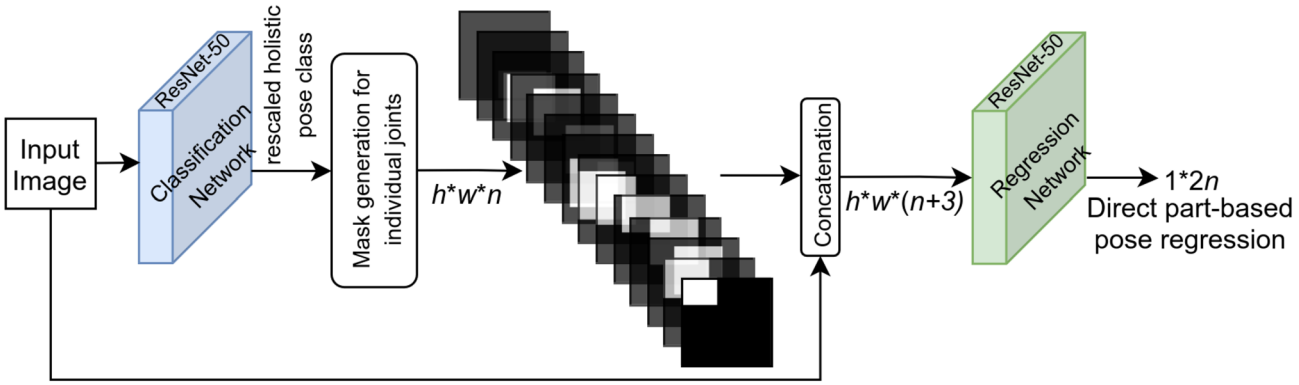

Fig. 5 The pipeline for the proposed series combination of the holistic and the part-based methods. h and w are the spatial dimensions of the input data and n indicates the number of joints

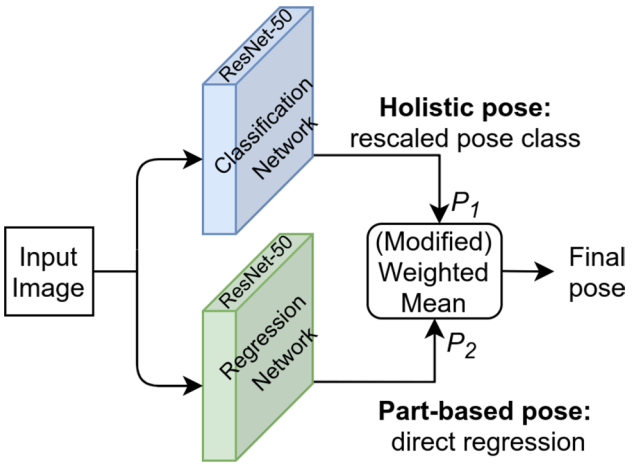

Fig. 6 The block diagram of the parallel combination of the holistic and the part-based pose estimations

model. In other words, not only the network searches the whole input image for a specific joint, but it concentrates more on the areas around that joint. Therefore, because of focusing on specific areas, we expect that the series method performs with robustness where the direct part-based human pose estimation fails and brings in false detection of the joints or invalid pose estimation, like in heavy occlusion of a joint, cluttered scenes, and in images with low quality or resolution.

3.5 Parallel fusion

In this technique, the holistic and the part-based networks are trained and tested independently. In other words, each network makes a prediction for the human pose and then, the two poses are combined with a combination formula. The block diagram of the parallel fusion is illustrated in Fig. 6. We propose two types for pose combination as follows:

- Weighted mean: To this end, the weighted mean of the corresponding joints is computed:

$$P = \lambda P_1 + (1 - \lambda) P_2, \tag{2}$$

where P_1 and P_2 are the computed poses using the holistic and the part-based networks, respectively, and λ represents a factor of combination. That is, the final location for i-th joint is:

$$\begin{aligned} x_i &= \lambda x_i^1 + (1 - \lambda) x_i^2 \\ y_i &= \lambda y_i^1 + (1 - \lambda) y_i^2, \end{aligned} \tag{3}$$

where (x_i^1, y_i^1) and (x_i^2, y_i^2) are the predicted 2D locations of the i-th joint using the holistic and the part-based methods, respectively. Since we expect the part-based method gets closer to the joint locations, the coefficient for the part-based method should be higher.
- Modified weighted mean: In this case, the corresponding joint distances between P_1 (holistic pose) and P_2 (part-based pose) are computed. Then, the holistic pose is shifted in a way that the joint with the least distance is placed in the same joint location of the part-based pose. After shifting, once again, the corresponding joint distances between the part-based pose and the shifted holistic pose are computed. The distances are normalized w.r.t the torso diagonal computed from the part-based method (P_2). If the normalized distance for a joint is less than a threshold, the weighted mean of P_1 and P_2 is considered for that joint, and if the distance is more than a threshold, only the result of P_2 is considered for the joint location. The reason for combining the two estimated poses with P_2 as the hub is that P_2 is expected to estimate finer joint locations (unlike the approximate and coarse joint detections of P_1).

3.6 Assisting the network

Here, we explain how it is possible to assist the feature learning in a deep neural network by directly incorporating a lower-level feature in the learning process. To this end, we add an extra channel along with the RGB image as the input to the network. Therefore, the input to the network would be of size $h \times w \times 4$. This supplemented image is considered to be an edge map and the motivation behind this supplementation is to make the network focus on the information that are useful for estimating the human pose, and ignore image data that can be deceptive for the goal. Accordingly, the edge probabilities of the human shape contour is a effective feature for human pose estimation, while color, texture, and small variations in the background can mislead the performance (see Fig. 7). Thus, by using a lower-level image feature, we encourage the network to learn the useful information for human pose prediction. As far as we know, this is the first time that a lower-level image feature is used directly as an input in the learning process of a deep network in order to gain focus towards the defined goal. The method of Dolláar and Zitnick (2015) is used for edge map extraction.

4 Results and discussion

In this section, the implementation details are explained and the suggested three fusion techniques are evaluated. Based on the obtained results, we then utilize both the *parallel* and the *multitasking* techniques together as our final approach for estimating the human pose (Fig. 1).

4.1 Implementation details

For implementation, we have used 19,882, 1000, and 3686 number of images from the training sets of MPII (Andriluka et al. 2014), LSP, and LSPextended (Johnson and Everingham 2011), respectively. The reason for choosing a lower number than the available ones in the MPII and the LSPextended is that for the holistic network, we need samples in which all the 14 joints, as in the LSP dataset, have been annotated. For augmentation, the images and their annotations are horizontally flipped, reaching to a number of 49,136 ones. Furthermore, we rotate these sample from −40° to 40° with step size of 10°. Consequently, the augmented dataset would have a total number of 442,224 samples.

We randomly divide the dataset into two subsets, $S1$ and $S2$. The reason for this division is twofold: (1) in the series scheme, we train the holistic and the part-based network separately. Therefore, two non-overlapping training sets are considered for each case (explained later), (2) we trained different scenarios with the first half of the dataset to analyze their overall performance and then, the best case is trained with all the training set (441 K samples). Note that a number of 1000 samples are selected randomly from $S1$ to act as a validation set. So, the number of training samples in $S1$ is 220,112. Finally, the test set in all the experiments are 1000 test samples of the LSP dataset.

We have used the network of ResNet-50 (He et al. 2016), pre-trained on the ImageNet dataset (Russakovsky et al. 2015), for our proposed frameworks. In our implementations, the network is trained with the stochastic gradient descent (SGD) optimization with momentum and with the learning rate of 10^{-3} for 30 epochs with batch size of 32. We have used MatConvNet (Vedaldi and Lenc 2015) and a system with two NVIDIA GeForce GTX 1050 for simulations.

Fig. 7 Edge maps of some sample images. This type of lower-level feature can provide helpful cues for human pose estimation and discard the misleading image feature, like color, texture, etc., in the model learning process

Table 1 PCK-0.2 (%) accuracy on the LSP test set using the holistic ("H") and the part-based ("P") methods

	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
H	58.95	59.90	47.50	34.25	64.80	43.85	33.00	48.89
H–E	58.95	59.86	47.05	34.20	64.65	43.57	33.85	48.87
P	94.90	91.35	83.50	76.35	90.80	88.90	85.40	87.31
P–E	95.05	91.10	83.85	77.40	90.00	89.95	87.00	**87.76**

The additional "E" states that the input image is supplemented with the edge map

Bold value shows the maximum value in each column

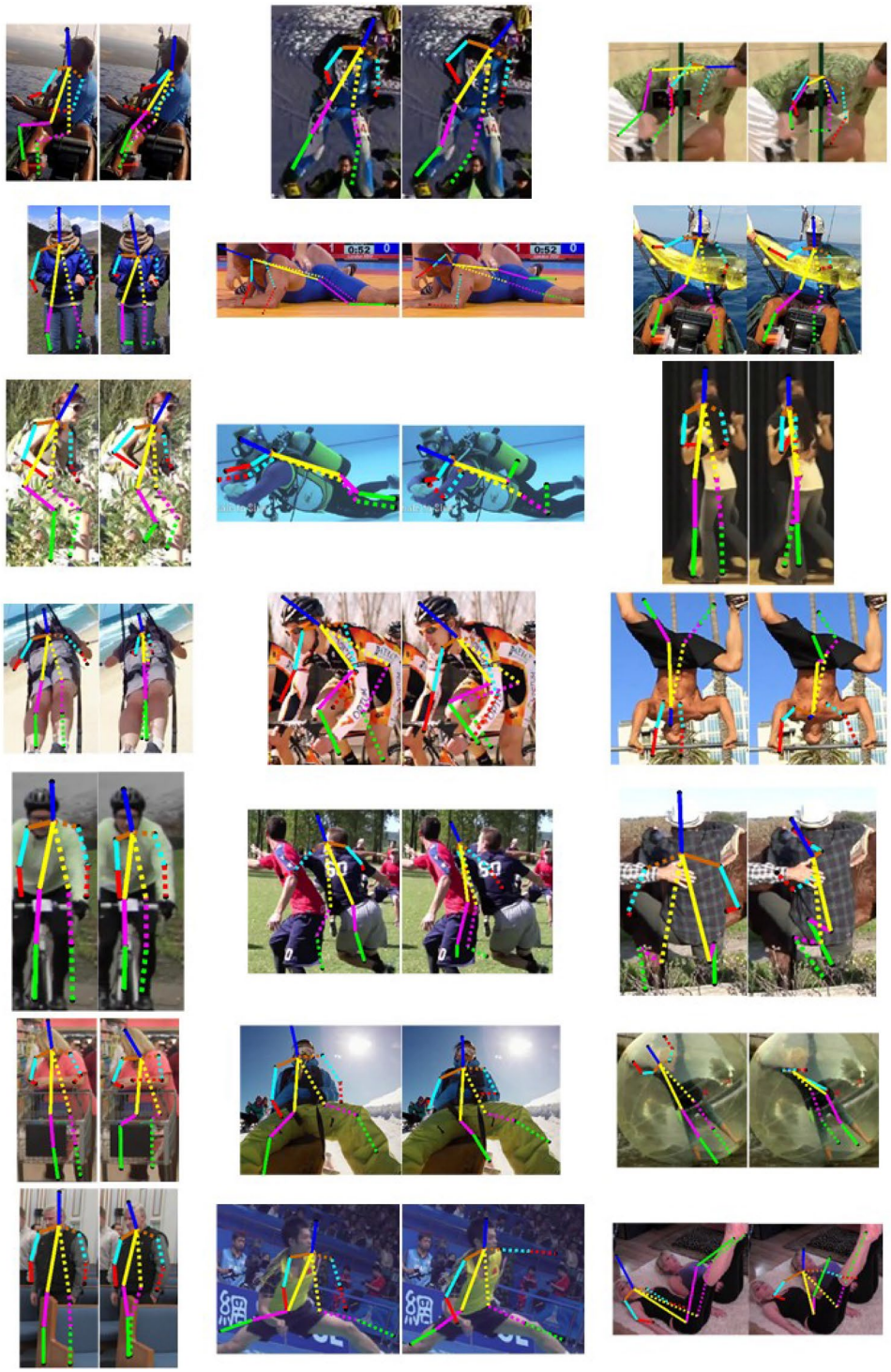

Fig. 8 Qualitative comparison between the holistic and the part-based approaches. In each pair, the left and the right images show the results of the holistic and the part-based methods in order. Note the robustness of the holistic method in the challenging joints

Fig. 9 Qualitative comparison between the part-based approach with and without using the edge map along with the input image. In each pair, the left and the right images show the results of the input data as RGB and RGB-E (RGB + edge map) in order. Incorporating the lower-level edge feature improves the performance of the network

In the classification (holistic) method, the procedure of assigning a ground truth pose class label to each sample is based on the least mean Euclidean distance to the representative of the pose class (Shamsafar and Ebrahimnezhad 2018). The number of pose classes is assumed to be 1000. This choice is not an optimum one; but the number of classes should be chosen such that the network can learn the similarity of images within one class.

The results of the holistic classification and the part-based regression are reported in Table 1. The mean PCK (Chen and Yuille 2014) for the holistic method is 48.89%, which is much lower than the one in the part-based

Table 2 PCK-0.2 (%) accuracy on the LSP test set using different combinations of the holistic and the part-based methods

	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
H[a]	58.95	59.90	47.50	34.25	64.80	43.85	33.00	48.89
P[b]	94.90	91.35	83.50	76.35	90.80	88.90	85.40	87.31
Multi-H	61.65	62.95	50.50	35.05	63.55	46.70	35.95	50.90
Multi-P	94.60	90.65	82.10	75.00	89.20	87.15	84.40	86.15
Series	94.75	91.22	83.45	76.20	**90.76**	88.72	85.35	87.20
Parallel	95.10	91.15	83.91	77.40	90.31	89.90	**87.12**	87.84
Parallel-M	**95.20**	**91.35**	**83.95**	**77.45**	90.50	**90.00**	86.85	**87.90**

"Multi-H" indicates the holistic output of the multitasking framework and "Multi-P" shows the part-based output. "Parallel-M" is the parallel combination with the modified weighted mean

Bold values show the maximum value in each column

[a]Holistic baseline, [b]Part-based baseline

method (87.31%). This difference was expected since the holistic method predicts a coarse pose in a classification manner, while the part-based method performs a regression task with numerous parameters in the mapping function. In other words, the holistic method pays attention to the overall similarity of the pose, whereas the part-based loss function penalizes the false estimation for individual joints. The best and the least recognition rates in the holistic design belong to the hip joints and ankles, respectively. Although these holistic estimations cannot catch up with the accuracy of the part-based method in terms of the PCK metric, the holistic method surpasses the part-based one in image samples with severe difficulties, which make the part-based method get stuck in undesired data. Figure 8 illustrates some of these samples. In this figure, we can see how the holistic method shows robustness in occlusion/self-occlusion (e.g. the image in row7/column1), low quality (e.g. the image in row6/column3), cluttered background/false detection (e.g. the image in row7/column3) and double counting (e.g. the image in row3/column3). The holistic model ("H" in Table 1) and the part-based approach ("P" in Table 1) are considered as the baselines of each type in our experiments.

Additionally, we investigate the effect of adding an edge map to the input data in order to assist the network in focusing on helpful image cues. In the holistic approach, adding the edge map does not make a difference on the mean PCK, see Table 1. This is caused by the intrinsic characteristic of the holistic approach, which does not pay attention to the details in the image. However, in the part-based approach, there are improvements in joints of the head, elbows, wrists, knees, ankles, and consequently, on the mean of all joints. The qualitative improvement by adding the edge map to the input data is shown in Fig. 9. Note how the direct addition of the lower-level edge feature, improves the prediction, especially in images with high articulation (e.g. the images in row1/column3, row3/column3, row6/column1 and row5/column3).

4.2 Multitasking fusion

In the multitasking framework, whose results are reported in Table 2, it is observed that the mean PCK for the part-based output of the network has decreased in comparison to the part-based baseline, whereas the same for the holistic pose has grown. This is due to the nature of the methods; the part-based method tries to predict the exact joint locations and it is more accurate than the holistic one and therefore, it is expected that by the joint training procedure, the performance of the part-based and the holistic approaches warp toward each other. In other words, while the holistic prediction enhances, the part-based strategy deteriorates. The average gain of accuracy in the holistic output of the multitasking network in comparison to the holistic baseline is 4.82%, with the highest gain in the ankles (8.93%). Thus, multitasking can be beneficial in tasks that require only a valid coarse understanding of the human pose (e.g. in content-based image retrieval, advanced driving assistance systems, patient monitoring), or in cases where there is memory shortage hindering us from using two types of holistic and part-based networks. In our final model, we will make use of the holistic output of the multitasking framework ("Multi-H" in Table 2).

4.3 Series fusion

In the series method, as stated previously, the holistic and the part-based methods are trained separately. The holistic network is fine-tuned on samples of the $S1$ set. After training the holistic network in a classification scheme, samples of the $S2$ set are fed as test samples. Now, the images of $S2$ along with the generated mask images from the holistic network are exploited for the part-based network. The pre-trained ResNet-50 maps the concatenated input data to a vector of size 1×28.

The result of the series approach is tabulated in Table 2. The mean PCK accuracy in the series fusion (87.20%) is lower than the part-based baseline (87.31%). This can be

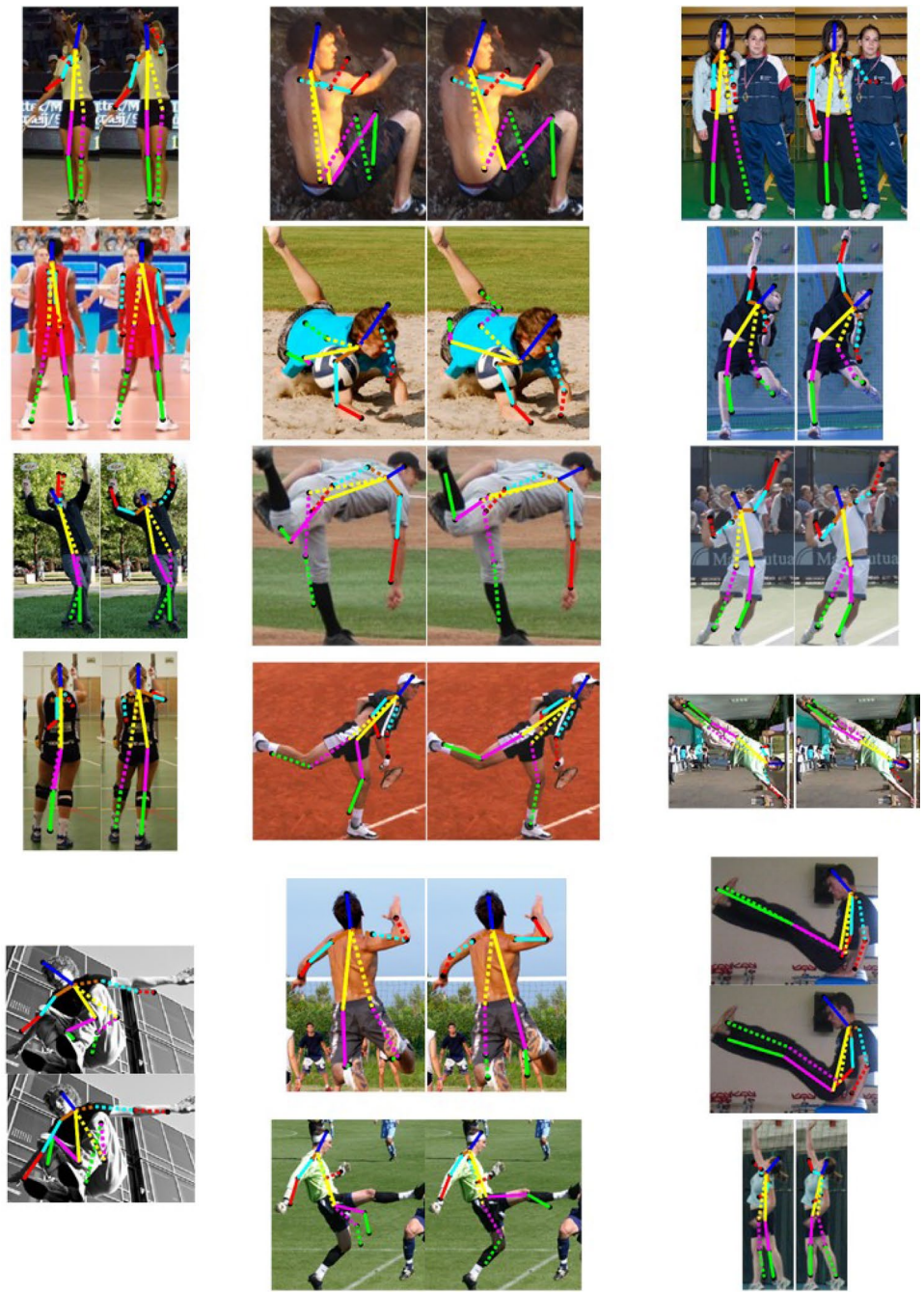

Fig. 10 Qualitative comparison between the part-based baseline and the series approach. In each pair, the left/top and right/bottom images show the results of the part-based and the series methods in order

originated from the fact that in the holistic prediction, some estimated joint locations are not close enough to the correct joint locations and accordingly, the corresponding joint mask images do not contain the correct joint. The series scenario has also been conducted using the Gaussian masks (instead of the square binary masks) and using two smaller and larger square sizes, both of which decreased the accuracy of estimation.

Although the series fusion of the holistic and the part-based methods has reduced the mean PCK a bit, it enhances the prediction in some samples, as shown in Fig. 10. In each pair, the left/top sample shows the poses estimated by the part-based baseline, and the right/bottom one shows prediction using the series pipeline. The part-based approach has confused the right/left body parts in some of these challenging images (e.g. the image in row4/column2). In some other images, predictions of the part-based method yield some erroneous joints, e.g. in the sample in row3/column1. These inaccuracies, even in the joints that are severely occluded (e.g. the image in row1/column3), have been solved to some extent by utilizing the series combination.

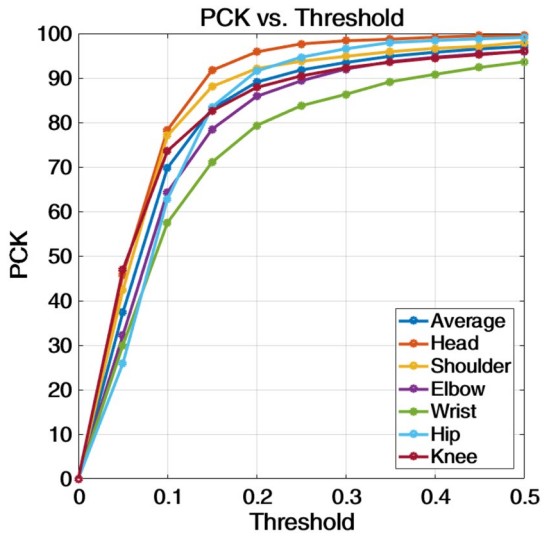

Fig. 11 PCK (%) for different joints on the LSP test set, with respect to varying amounts of threshold. The approach is part-based with the edge map ("P–E") and trained on about 440 K training samples

Hence, one can observe that in all these problematic images, the series scheme is capable of predicting much closer locations to the ground truth joints by considering the neighborhood of the joints. While this strategy shows more robustness than the part-based baseline in the challenging real-world images, yet, its estimations cannot pass the threshold for being considered as a correct prediction. Note that in the PCK metric, the accuracy is decided based on a threshold. This means that maybe a prediction is much closer to the ground truth joint, but it still cannot pass the threshold in this type of strict binary decision, and cannot go from the false detection zone to the true detection region. This is another reason why we have not obtained a higher PCK value. It seems the PCK metric has not enough flexibility to truly evaluate how much improvement has been made by an algorithm in estimating the human pose.

4.4 Parallel fusion

The results of the parallel combination using the weighted mean and the modified weighted mean are shown in Table 2. Note that we have used part-based method along with the edge map input, i.e. "P–E", for this fusion since this supplementation improves the part-based estimation.

Both types of the parallel combination have increased the mean PCK accuracy. This implies that the holistic prediction can improve the results of part-based method independently. Having a closer evaluation of the results, we found that the parallel fusion can relocate one or two joint predictions per sample, such that they can pass the threshold of the PCK metric and this leads to more correct detection amount.

When comparing the weighted mean combination with the modified one, the latter performs better. This shows that the computed coarse holistic pose can get closer to the correct pose by an amount of translation. The modified weighted mean case works better in joints of head, shoulders, hips and knees.

4.5 Training with all the dataset

Here, we train the network that demonstrates the best accuracy from the previous section, i.e. the part-based method with the edge map input ("P–E"), with all the dataset. In previous experiments, only half of the augmented dataset ($S1$) was used for training the network. The computed accuracy rates are listed in Table 2. The mean PCK has risen from 87.7 to 89.05%. Figure 11 shows the mean PCK for different joints using the part-based approach with the edge map trained on about 440 K images. According to this figure, the best accuracy belongs to the head joints while the wrists show the least recognition rate. Wrist joints are the most difficult joints to detect because of the high variability of their location and their small-size and low-resolution appearance.

As a final approach in our experiments, we have parallel-combined the methods of the part-based with the edge map ("P–E") together with the holistic pose computed from the multitasking network ("Multi-H") using the modified weighted mean function. The related block diagram is the one shown previously in Fig. 1. By this type of combination, we can reach to a higher mean recognition accuracy, 90.01% (Table 3).

Figure 12 shows the results of our final framework on some samples from the LSP test set. We can see that in various cases of challenges of pose estimation, namely high articulation, high appearance variability (color and texture of clothes, body shape, skin color) and low quality,

Table 3 PCK-0.2 (%) on the LSP test set: "P–E" indicates the part-based method that is assisted by the edge map information

	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
P–E	95.85	92.10	85.85	79.30	91.55	90.85	87.85	89.05
Parallel-M	97.14	93.73	86.97	79.30	93.08	91.99	87.89	90.01

The network is trained with all the augmented dataset. "Parallel-M" is the parallel combination of "P–E" and the holistic pose obtained from the multitasking framework in Sect. 4.2

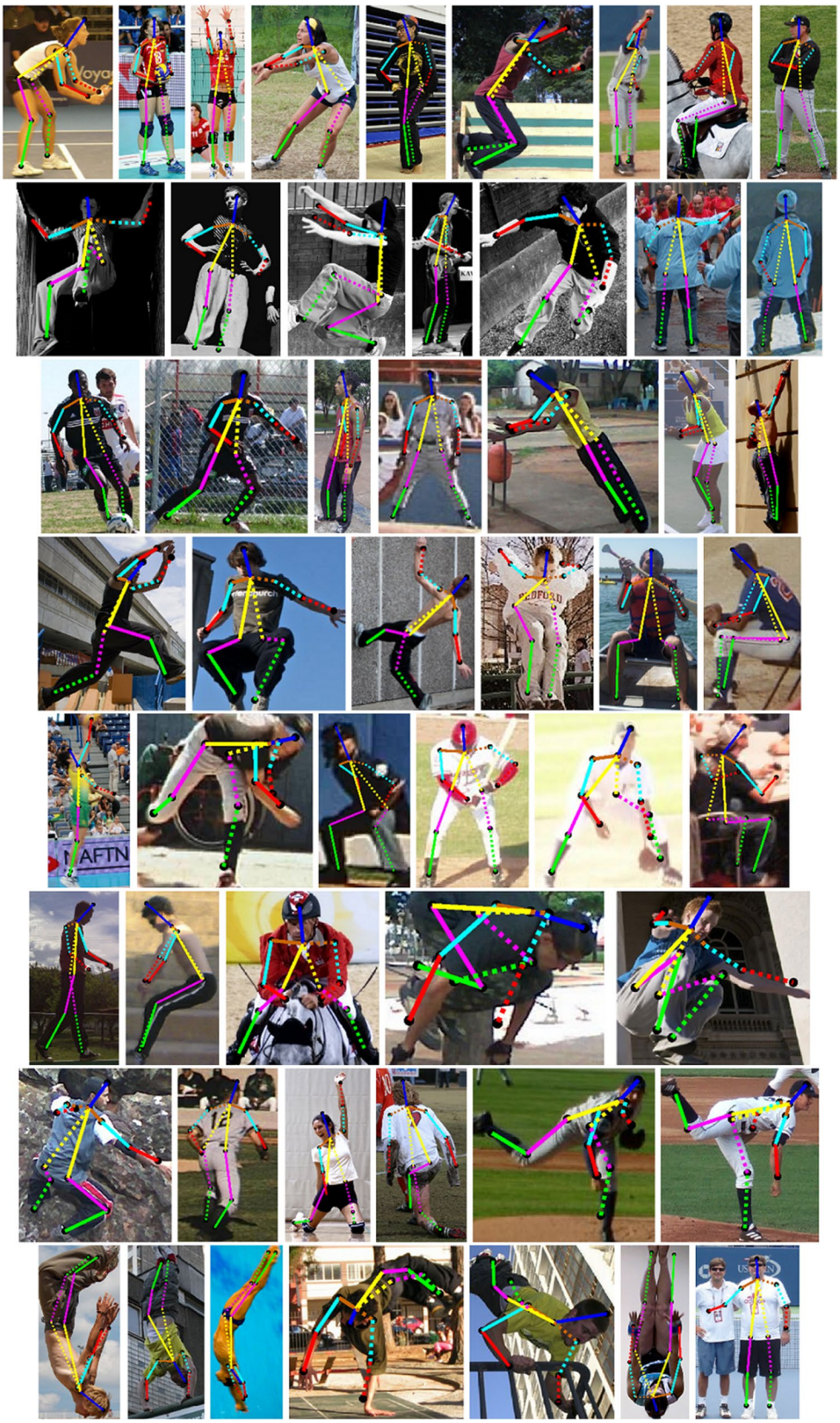

Fig. 12 The result of our proposed final model in challenging images. Note how the model has performed robust in high articulation, low quality, light variation, heavy occlusion and cluttered background

Table 4 Comparisons of the PCK-0.2 (%) on the LSP test set

	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Wang and Li (2013)	84.7	57.1	43.7	36.7	56.7	52.4	50.8	54.6
Pishchulin et al. (2013b)	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1
Tompson et al. (2014)	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3
Fan et al. (2015)	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0
Carreira et al. (2016)	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1
Yang et al. (2016)	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6
Ukita and Uematsu (2018)	93.6	85.1	76.3	71.0	85.2	80.6	77.8	81.4
Rafi et al. (2016)	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8
Yu et al. (2016)	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Belagiannis and Zisserman (2017)	95.2	88.7	81.7	76.8	83.8	86.7	82.5	85.1
Lifshitz et al. (2016)	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin et al. (2016)	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Our model	**97.1**	**93.7**	**86.9**	**79.3**	**93.0**	**91.9**	**87.8**	**90.0**

Bold values show the maximum value in each column

Table 5 Comparisons of the PCP-0.5 (%) on the LSP test set

	U-arms	L-arms	U-legs	L-legs	Head	Torso	Mean
Fan et al. (2015)	56.0	38.0	77.0	71.0	–	–	–
Wang and Li (2013)	43.1	32.1	56.0	55.8	79.1	87.5	54.1
Pishchulin et al. (2013b)	46.0	35.2	63.6	58.4	85.1	88.7	58.0
Kiefel and Gehler (2014)	54.1	28.3	74.5	67.6	78.3	85.8	61.2
Yang and Ramanan (2013)	56.0	39.8	70.3	67.0	79.3	88.7	62.8
Pishchulin et al. (2013a)	54.2	33.9	75.7	68.0	78.1	87.5	62.9
Eichner and Ferrari (2012)	56.5	37.4	74.3	69.3	80.1	86.2	64.3
Tompson et al. (2014)	63.0	51.2	70.4	61.1	83.7	90.3	66.6
Ramakrishna et al. (2014)	62.8	39.5	79.0	73.6	80.4	88.1	67.8
Ouyang et al. (2014)	63.3	46.6	76.5	72.2	83.1	84.3	68.6
Hernández-Vela et al. (2016)	57.6	42.0	77.3	72.9	84.2	88.4	67.2
Belagiannis et al. (2015)	61.3	40.3	79.9	74.3	83.2	92.7	68.8
Pishchulin et al. (2012)	61.8	45.0	78.9	73.2	85.1	82.9	69.2
Fan et al. (2015)	62.8	49.1	77.7	69.8	86.6	95.4	70.1
Carreira et al. (2016)	66.7	51.0	81.8	73.3	84.4	95.3	72.5
Chen and Yuille (2014)	69.7	58.1	77.2	72.2	85.6	96.0	73.6
Chu et al. (2016)	76.0	64.3	87.6	83.5	89.4	95.4	80.8
Yang et al. (2016)	66.7	78.8	88.7	81.7	83.1	96.5	81.1
Rafi et al. (2016)	76.8	66.2	87.3	80.2	93.3	97.6	81.2
Belagiannis and Zisserman (2017)	79.4	69.4	86.7	82.2	89.4	96.0	82.1
Lifshitz et al. (2016)	80.4	71.4	88.9	84.5	94.7	97.3	84.2
Pishchulin et al. (2016)	82.4	71.8	88.8	82.0	95.8	97.0	84.3
Yu et al. (2016)	82.9	72.6	93.1	88.1	83.0	**98.0**	85.4
Our model	**85.6**	**73.2**	**94.0**	**91.4**	**92.1**	95.3	**87.6**

The prefixes "U-" and "L-" indicate the upper and the lower parts in order (e.g. U-arms is the upper arms)
Bold values show the maximum value in each column

the method achieves robust and accurate predictions. In addition, it does not get stuck in false part detections when similar body parts exist in the cluttered background. Body part occlusion/self-occlusion is another major problem of human pose estimation in the wild, particularly in 2D images when the amount of available cues is limited. Our method is able to correctly predict the occluded body parts even in totally occluded ones. As illustrated in Fig. 12, the performance also demonstrate robustness when the viewing angle is different.

We compare our final proposed approach with some works based on the PCK-0.2 and the PCP-0.5 (Eichner and Ferrari 2009; Felzenszwalb and Huttenlocher 2005) metrics in Tables 4 and 5 in order. Our proposed method obtains higher recognition rate in terms of both PCK-0.2 and PCP-0.5 values (except in torso). The suggested paradigm for human pose estimation has a feed-forward procedure and unlike other work, it does not need feedback or repetition.

As discussed earlier, popular types of evaluation metrics for pose estimation, e.g. PCK-0.2 and PCP-0.5, strictly decide based on one threshold, and they can not convey fairly how one method performs better in comparison to the other one. Also, there is room to measure the robustness of the performance by a metric. This issue is gaining much importance since the computer vision and the deep learning communities are moving towards applying the algorithms in the wild, where there is no constraint. In our future work, we aim to introduce an evaluation metric, which can act softer and also assess the robustness of an algorithm in addition to its accuracy.

5 Conclusion

In this paper, we proposed to estimate the human pose simultaneously based on two viewpoints: holistic and part-based predictions. The holistic framework executes a pose classification task in a deep network, whereas the part-based prediction requires a regression mapping. The motivation behind this fusion is as follows: the part-based method predicts more accurate results, but gets stuck in challenges of human pose estimation; on the other hand, the holistic mode promotes robustness in severe challenges, while it does not get close enough to joint locations. These two frameworks were combined to compensate the weakness of each method.

In practice, the multitasking network boosts the accuracy of the holistic approach; the series network enhances the robustness of pose estimation in challenges that cause false detections when using the part-based method, and lastly, in the parallel scheme, the mean recognition rate is improved directly. Based on the outcomes obtained from the conducted experiments, we fused the *parallel* and *multitasking* schemes as our final proposed model. The results demonstrate that for more accurate and for more robust human pose estimation, the visual information of human beings should be processed both holistically and based on individual parts.

Compliance with ethical standards

Conflict of interest Faranak Shamsafar declares that she has no conflict of interest. Hossein Ebrahimnezhad declares that he has no conflict of interest.

Funding This research received no specific grant from any funding agency in the public, commercial, or non-profit sectors.

References

Agarwal A, Triggs B (2006) Recovering 3D human pose from monocular images. IEEE Trans Pattern Anal Mach Intell 28(1):44–58. https://doi.org/10.1109/TPAMI.2006.21

Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2D human pose estimation: new benchmark and state of the art analysis. In: IEEE conference on computer vision and pattern, pp 3686–3693. https://doi.org/10.1109/CVPR.2014.471

Belagiannis V, Rupprecht C, Carneiro G, Navab N (2015) Robust optimization for deep regression. In: International conference on computer vision, pp 2830–2838. https://doi.org/10.1109/ICCV.2015.324

Belagiannis V, Zisserman A (2017) Recurrent human pose estimation. In: IEEE international conference on automatic face and gesture recognition, pp 468–475. https://doi.org/10.1109/FG.2017.64

Carreira J, Agrawal P, Fragkiadaki K, Malik J (2016) Human pose estimation with iterative error feedback. In: IEEE conference on computer vision and pattern, pp 4733–4742. https://doi.org/10.1109/CVPR.2016.512

Chen X, Yuille A (2014) Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in neural information processing systems, pp 1736–1744

Chu X, Ouyang W, Li H, Wang X (2016) Structured feature learning for pose estimation. In: IEEE conference on computer vision and pattern, vol 2016-Dec, pp 4715–4723. https://doi.org/10.1109/CVPR.2016.510. arXiv:1603.09065

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. IEEE Conf Comput Vis Pattern 1:886–893

Dolláar P, Zitnick CL (2015) Fast edge detection using structured forests. IEEE Trans Pattern Anal Mach Intell 37(8):1558–1570. https://doi.org/10.1109/TPAMI.2014.2377715

Eichner M, Ferrari V (2009) better appearance models for pictorial structures. In: British machine vision conference, pp 3.1–3.11. DOIurlhttps://doi.org/10.5244/C.23.3.arXiv:1504.08083

Eichner M, Ferrari V (2012) Appearance sharing for collective human pose estimation. In: Asian conference on computer vision. Springer, Berlin, pp 138–151

Fan X, Zheng K, Lin Y, Song W (2015) Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation. In: IEEE conference on computer vision and pattern, pp 1347–1355. https://doi.org/10.1109/CVPR.2015.7298740

Felzenszwalb PF, Girshick RB, McAllester D (2010a) Cascade object detection with deformable part models. In: IEEE conference on computer vision and pattern, pp 2241–2248. https://doi.org/10.1109/CVPR.2010.5539906

Felzenszwalb PF, Huttenlocher DP (2005) Pictorial structures for object recognition. Int J Comput Vis 61(1):55–79

Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010b) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645

Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multi-scale, deformable part model. In: IEEE conference on computer vision and pattern, pp 1–8

Gavrila DM (2007) A Bayesian, exemplar-based approach to hierarchical shape matching. IEEE Trans Pattern Anal Mach Intell 29(8):1408–1421. https://doi.org/10.1109/TPAMI.2007.1062

Hernández-Vela A, Sclaroff S, Escalera S (2016) Poselet-based contextual rescoring for human pose estimation via pictorial structures. Int J Comput Vis 118(1):49–64

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern, pp 770–778. https://doi.org/10.1109/CVPR.2016.90

Jain A, Tompson J, Andriluka M, Taylor GW, Bregler C (2014) Learning human pose estimation features with convolutional networks. In: International conference on learning representations. arXiv:1312.7302

Johnson S, Everingham M (2010) Clustered pose and nonlinear appearance models for human pose estimation. In: British machine vision conference, pp 12.1–12.11. https://doi.org/10.5244/C.24.12

Johnson S, Everingham M (2011) Learning effective human pose estimation from inaccurate annotation. In: IEEE conference on computer vision and pattern, pp 1465–1472. https://doi.org/10.1109/CVPR.2011.5995318

Kiefel M, Gehler PV (2014) Human pose estimation with fields of parts. In: European conference on computer vision, pp 331–346

Kokkinos I (2012) bounding part scores for rapid detection with deformable part models. In: European conference on computer vision, vol 7585 LNCS, pp 41–50

Li S, Liu ZQ, Chan AB (2015) Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. Int J Comput Vis 113(1):19–36. https://doi.org/10.1007/s11263-014-0767-8. arXiv:1406.3474

Lifshitz I, Fetaya E, Ullman S (2016) Human pose estimation using deep consensus voting. In: European conference on computer vision, pp 246–260

Liu T, Liu J, Xm Luo (2014) Radio tomographic imaging based body pose sensing for fall detection. J Ambient Intell Humaniz Comput 5(6):897–907

Mori G, Malik J (2002) Estimating human body configurations using shape context matching. In: European conference on computer vision, pp 666–680. https://doi.org/10.1007/3-540-47977-5

Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. Int Conf Pattern Recogn 1:582–585. https://doi.org/10.1109/ICPR.1994.576366

Ouyang W, Chu X, Wang X (2014) Multi-source deep learning for human pose estimation. In: IEEE conference on computer vision and pattern, pp 2329–2336

Pishchulin L, Andriluka M, Gehler P, Schiele B (2013a) Poselet conditioned pictorial structures. In: IEEE conference on computer vision and pattern, pp 588–595. https://doi.org/10.1109/CVPR.2013.82

Pishchulin L, Andriluka M, Gehler P, Schiele B (2013b) Strong appearance and expressive spatial models for human pose estimation. In: International conference on computer vision, pp 3487–3494

Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler P, Schiele B (2016) DeepCut: joint subset partition and labeling for multi person pose estimation. In: IEEE conference on computer vision and pattern, pp 4929–4937. https://doi.org/10.1109/CVPR.2016.533

Pishchulin L, Jain A, Andriluka M, Thormählen T, Schiele B (2012) Articulated people detection and pose estimation: reshaping the future. In: IEEE Conference on computer vision and pattern, pp 3178–3185

Rafi U, Leibe B, Gall J, Kostrikov I (2016) An efficient convolutional network for human pose estimation. In: British machine vision conference, pp 109.1–109.11. https://doi.org/10.5244/C.30.109

Ramakrishna V, Munoz D, Hebert M, Andrew Bagnell J, Sheikh Y (2014) Pose machines: articulated pose estimation via inference machines. In: European conference on computer vision, vol 8690 LNCS, pp 33–47

Rogez G, Rihan J, Ramalingam S, Orrite C, Torr PH (2008) Randomized trees for human pose detection. In: IEEE conference on computer vision and pattern, pp 1–8. https://doi.org/10.1109/CVPR.2008.4587617

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252

Shakhnarovich G, Viola P, Darrell T (2003) Fast pose estimation with parameter-sensitive hashing. In: International conference on computer vision, pp 750–757 vol. 2. https://doi.org/10.1109/ICCV.2003.1238424

Shamsafar F, Ebrahimnezhad H (2018) Understanding holistic human pose using class-specific convolutional neural network. Multimed Tools Appl 77(18):23193–23225. https://doi.org/10.1007/s11042-018-5617-1

Sun X, Shang J, Liang S, Wei Y (2017) Compositional human pose regression. In: International conference on computer vision, pp 2621–2630. https://doi.org/10.1109/ICCV.2017.284. arXiv:1704.00159

Tompson J, Jain A, LeCun Y, Bregler C (2014) Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems, pp 1799–1807

Toshev A, Szegedy C (2014) DeepPose: human pose estimation via deep neural networks. In: IEEE conference on computer vision and pattern, pp 1653–1660. https://doi.org/10.1109/CVPR.2014.214

Ukita N, Uematsu Y (2018) Semi-and weakly-supervised human pose estimation. Comput Vis Image Underst 170:67–78

Vedaldi A, Lenc K (2015) MatConvNet: convolutional neural networks for MATLAB. In: ACM international conference on multimedia, pp 689–692. https://doi.org/10.1145/2733373.2807412. http://www.vlfeat.org/matconvnet/

Wang F, Li Y (2013) beyond physical connections: tree models in human pose estimation. In: IEEE conference on computer vision and pattern, pp 596–603

Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: IEEE conference on computer vision and pattern, pp 4724–4732. https://doi.org/10.1109/CVPR.2016.511

Yan C, Gong B, Wei Y, Gao Y (2020a) Deep multi-view enhancement hashing for image retrieval. IEEE Trans Pattern Anal Mach Intell 20:20

Yan C, Shao B, Zhao H, Ning R, Zhang Y, Xu F (2020b) 3d room layout estimation from a single RGB image. IEEE Trans Multimed 20:20

Yang Y, Ramanan D (2013) Articulated human detection with flexible mixtures of parts. IEEE Trans Pattern Anal Mach Intell 32(12):2878–2890

Yang W, Ouyang W, Li H, Wang X (2016) End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: IEEE conference on computer vision and pattern, pp 3073–3082

Yu X, Zhou F, Chandraker M (2016) Deep deformation network for object landmark localization. In: European conference on computer vision, vol 9909 LNCS, pp 52–70. arXiv:1605.01014

Zavala-Mondragon LA, Lamichhane B, Zhang L, de Haan G (2019) CNN-skelpose: a CNN-based skeleton estimation algorithm for clinical applications. J Ambient Intell Human Comput 20:1–12

Zhou X, Sun X, Zhang W, Liang S, Wei Y (2016) Deep kinematic pose regression. In: European conference on computer vision workshop, vol 9915 LNCS, pp 186–201. arXiv:1609.05317