Active Learning for Human Pose Estimation based on Temporal Pose Continuity

Taro Mori*[a], Daisuke Deguchi[a], Yasutomo Kawanishi[b,a]
Ichiro Ide[a], Hiroshi Murase[a], Tetsuo Inoshita[c]
[a] Graduate School of Informatics, Nagoya University, Nagoya, Japan
[b] RIKEN Information R&D and Strategy Headquarters, GRP, Kyoto, Japan
[c] NEC Corporation, Biometrics Research Laboratories, Kawasaki, Japan

ABSTRACT

In recent years, human pose estimation based on deep learning has been actively studied for various applications. A large amount of training data is required to achieve good performance, but, annotating human poses is quite an expensive task. Therefore, there is a growing need to improve the efficiency of training data preparation. In this paper, we take an active learning approach to reduce the cost of preparing training data for human pose estimation. We propose an active learning method that automatically selects images effective for improving the performance of a human pose estimation model from unlabeled image sequences, focusing on the fact that the human pose continuously changes between adjacent frames in an image sequence. Specifically, by comparing the estimated human poses between frames, we select images incorrectly estimated as candidates for manual annotation. Then, the human pose estimation model is re-trained by adding a small portion of manually annotated data as training data. Through experiments, we confirm that the proposed method can effectively select training data candidates from unlabeled image sequences, and that the proposed method can improve the performance of the model with reducing the cost of manual annotations.

Keywords: human pose estimation, active learning, deep learning

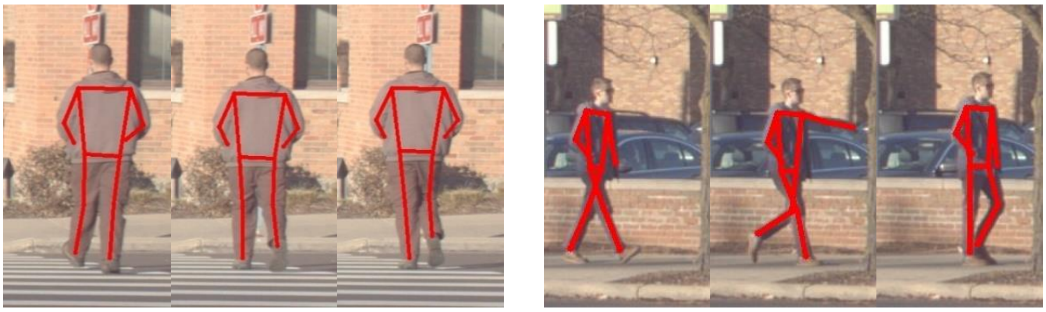
1. INTRODUCTION

In recent years, human pose estimation methods based on deep learning[1, 2, 3] have been widely studied, and their applications are being considered in various forms due to their high performance. As a typical application, action recognition based on human pose has been actively studied[4, 5]. Most of these applications require accurate pose estimation but a large amount of training data annotated with human poses should be prepared to construct high-performance deep learning-based pose estimation models. Since the training data requires accurate multiple joint positions, this annotation task is very time consuming. Thus, there is a strong demand for efficient training data preparation techniques.

Active learning is one of the most efficient approaches to reduce the cost of preparing training data[6, 7]. The approach expands the training data by automatically selecting and annotating images from a set of unlabeled images. Since the images are selected so that they can contribute to the performance improvement of a given model, it would be possible to build an accurate model with a lower cost.

B. Liu et al.[7] proposed an active learning method that automatically selects images in which the estimated human poses are uncertain, and annotates only those images to improve the accuracy with low annotation cost. The method uses not only the estimated human pose, but also the heat map estimated in the previous step of the estimation to evaluate the uncertainty. Specifically, the method focuses on the heat map of each human joint and determines uncertainty when there are multiple maxima in the heat map of each joint, i.e., when there are multiple candidates for that joint. Since this requires to estimate the heat map of human joints individually, this method can only be applied to a top-down method[3] that estimates the heat map of an individual human joint based on pedestrian detection technique. Therefore, it cannot be applied to a bottom-up method[1, 2] which estimates the heat map of all human joints simultaneously. From this point of view, in this paper, we propose an active learning method that can be applied to bottom-up human pose estimation methods; The proposed method automatically selects images from unlabeled image sequences that can contribute to the performance improvement if annotated.

*morit@vislab.is.i.nagoya-u.ac.jp



(a) $t-1$ frame (b) t frame (c) $t+1$ frame (d) $t-1$ frame (e) t frame (f) $t+1$ frame

Figure 1. (a)-(c) Examples of correctly estimated human poses in successive frames. The poses change smoothly between frames. (d)-(f) Examples of incorrectly estimated human poses. The left elbow and the wrist of the person in frame t (e) are incorrectly estimated, and their positions are significantly different from those of adjacent frames.

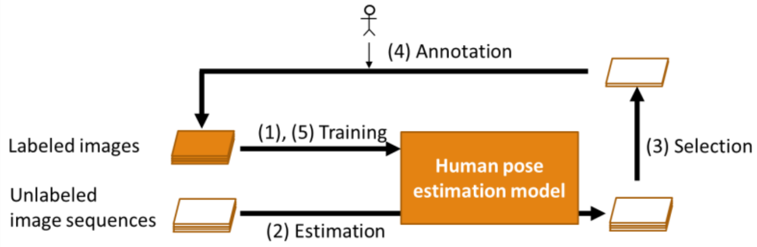


Figure 2. Process flow of the proposed method.

2. ACTIVE LEARNING USING IMAGE SEQUENCES

2.1 Overview of the proposed method

A careful observation of an image sequence that captures a walking pedestrian shows that the position of each human joint moves smoothly and continuously between adjacent frames. As shown in Figure 1, if the model can correctly estimate the human pose in a frame, the difference of the estimated human poses should be small between adjacent frames ((a)-(c)). On the contrary, if an estimated human pose has errors, the difference would be large ((d)-(f)). Furthermore, we notice a falsely undetected joint if the model detects the joint in both the previous and the next frames but not in the current frame. Based on these findings, we judge that a human pose in a frame is likely to be mis-estimated, if the estimated poses in the adjacent frames differ significantly with each other, or if the presence/absence of the estimated joints differs between the adjacent frames. The proposed method selects such images as annotation candidates for active learning.

The process flow of the proposed method is shown in Figure 2. The proposed method consists of five steps: (1) Initial training of a human pose estimation model using an existing labeled image dataset, (2) Estimation of human poses from unlabeled image sequences using the trained model, (3) Automatic selection of annotation candidates based on the estimation results, (4) Manual annotation of a portion of the candidates, and (5) Retraining of the human pose estimation model using existing labeled images and additional annotated images. In the following sections, steps (1) and (2) are explained in Section 2.2, a detailed explanation of step (3) is given in Section 2.3, and that of steps (4) and (5) are given in Section 0.

2.2 Human pose estimation from unlabeled image sequences

First, we construct an initial human pose estimation model by using an existing labeled image dataset. Next, we estimate human poses for unlabeled image sequences, using the constructed initial human pose estimation model. After that, we track the estimated poses between frames of the image sequences to obtain the time-series of individual human joints. We use PoseFlow[8], a human pose tracking method, to track the human poses.

2.3 Automatic selection of potentially mis-estimated images from unlabeled image sequences

In this step, the proposed method automatically selects images from the unlabeled image sequences that may have incorrectly estimated human joints. Then, those images are selected as candidates for manual annotation. Specifically, the proposed method calculates the likelihood of how much the model incorrectly estimates the human pose. Then, a portion of the candidates to be manually annotated are automatically selected in descending order of their likelihood.

For simplicity, we consider a single unlabeled image sequence. Here, $t \in \{1, \dots, T\}$ is a frame ID in an image sequence, $p \in \{1, \dots, P\}$ is a person ID and the coordinates of joint $j \in \{1, \dots, J\}$ is described as $y_{p,j}^{t} = \left(u_{p,j}^{t}, v_{p,j}^{t}\right)$. $e_{p,j}^{t} \in \{0,1\}$ indicates whether the joint is detected or not. If $e_{p,j}^{t} = 0$, the proposed method considers $y_{p,j}^{t} = (0,0)$. $\mathbb{D}_{p}^{t} = \left\{j | e_{p,j}^{t} = 1\right\}$ represents the set of estimated joints of person p at frame t.

First, from the estimated human pose, the proposed method obtains size S_p of each person in the unlabeled image sequences. Since the presence or absence of estimated human joints varies between frames, S_p is defined based on the area of the rectangle surrounding the largest pose in the image sequence, and is calculated as

$$S_p = \max_{t} \left\{ \left(\max_{j \in \mathbb{D}_p^t} \left(u_{p,j}^{t}\right) - \min_{j \in \mathbb{D}_p^t} \left(u_{p,j}^{t}\right) \right) \left(\max_{j \in \mathbb{D}_p^t} \left(v_{p,j}^{t}\right) - \min_{j \in \mathbb{D}_p^t} \left(v_{p,j}^{t}\right) \right) \right\}. \tag{1}$$

Next, the proposed method calculates the likelihood of how much the model incorrectly estimates the human pose using the estimated poses at frames $t - 1, t, t + 1$. Likelihood C^t is obtained by the following steps:

First, likelihood C_L^t is calculated based on the differences of the estimated joint positions. If the joint positions at frame t is incorrectly estimated, its position will differ significantly from those of frames $t - 1$ and $t + 1$. Based on this idea, we obtain C_L^t based on the difference in joint positions. Specifically, by using the threshold θ that is adjusted based on the size of the person, we count the number of incorrectly estimated joints $L_p^{t-1,t}$ whose Euclidean distance between frames $t - 1$ and t is greater than the threshold.

$$L_p^{t-1,t} = \sum_{j} e_{p,j}^{t-1} e_{p,j}^{t} \, \mathbf{1}\left[y_{p,j}^{t-1}, y_{p,j}^{t}\right], \tag{2}$$

where $\mathbf{1}[y_{p,j}^{t-1}, y_{p,j}^{t}]$ is given by

$$\mathbf{1}\left[y_{p,j}^{t-1}, y_{p,j}^{t}\right] = \begin{cases} 1 & \left\|y_{p,j}^{t-1} - y_{p,j}^{t}\right\| > \theta \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

and θ is adjusted as

$$\theta = \alpha \sqrt{S_p}\,, \tag{4}$$

where α is a hyperparameter, and $\alpha = 0.01$ is used in the experiments. C_L^t is calculated by summing $L_p^{t-1,t}$ and $L_p^{t,t+1}$ as

$$C_L^t = \sum_{p} \left(L_p^{t-1,t} + L_p^{t,t+1}\right). \tag{5}$$

Next, Likelihood C_U^t is calculated based on whether a joint is detected or not. When we focus on a certain joint of a person, the joint at frame t can be assumed as undetected, if the joint estimated at frames $t - 1$ and $t + 1$ is not estimated at frame t. Based on this idea, C_L^t is obtained as

$$C_U^t = \sum_{p,j} e_{p,j}^{t-1} \, e_{p,j}^{t+1} \left(1 - e_{p,j}^{t}\right). \tag{6}$$

Finally, Likelihood C^t is calculated by summing the likelihoods C_L^t corresponding to the difference of joint positions and C_U^t corresponding to undetected joints as

$$C^t = C_L^t + C_U^t\,. \tag{7}$$

This step is applied to all unlabeled image sequences frame-by-frame.

2.4 Re-training of the human pose estimation model

We automatically select candidates to be annotated from unlabeled image sequences and manually annotate them.

Table 1. Breakdown of the dataset

Training set		**Test set**
Labeled	Unlabeled	Labeled
56,599	3,188	1,971

Specifically, we use the unlabeled image sequences as input and sort them in descending order by the likelihood C^t calculated in Section 2.3 and select a portion of the candidates by descending order of likelihood for annotation. Then, the human pose estimation model is retrained using the newly annotated images and the existing training data.

3. EVALUATION

We conducted experiments to evaluate the effectiveness of the proposed method using publicly available datasets. Section 3.1 describes in detail the used datasets, Section 3.2 describes the experimental methods, and Section 3.3 describes the results.

3.1 Datasets

In the experiments, we used Microsoft COCO dataset[9] and PedX dataset[10]. The former was used for training the initial human pose estimation model, while the latter was used for extracting annotation candidates by the proposed method, and used for retraining the human pose estimation model. Table 1 shows the details of these two datasets. We extracted 56,599 images from the Microsoft COCO training set, and used as labeled images in the initial training step. PedX dataset contains labeled and unlabeled images, but we used only the labeled images for the evaluation. In the experiment, we divided the labeled images of the PedX dataset into two groups: (i) 3,188 images and (ii) 1,971 images. Here, (i) is used as unlabeled image sequences for re-training of the model, and (ii) is used for evaluation. Here, we split the PedX dataset so that neither the same image nor the same person is included in the training and evaluation datasets. Although the training dataset constructed from the PedX dataset has pose annotations, these images are used as unlabeled image sequences and the annotations are used instead of manual annotations in the retraining step of the proposed method. In this experiment, we used the following 12 joints: left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, and right ankle.

3.2 Experimental methods

First, we constructed an initial human pose estimation model trained by labeled images from the Microsoft COCO dataset. Then, this was used to estimate the human pose from the unlabeled image sequences, and we tracked the estimated poses in the image sequence to obtain the time-series of individual estimated poses. If the sum of the pose distances from the previous and the next frames was greater than a threshold θ, the estimated pose is not taken into account in the calculation of likelihood C^t to reduce the effect of tracking failure. Then, the likelihood was calculated for each unlabeled image sequence following the steps explained in Section 2.3. Next, we automatically selected the candidates for annotation from the unlabeled image sequences by referring to likelihood C^t, and annotated them. For images with equal likelihood, we randomly selected the candidates. In this experiment, instead of manually annotating the images, we used the human pose annotations from the PedX dataset. Finally, we re-trained the human pose estimation model using the training data with newly annotated images, and evaluated its performance using the evaluation data.

We used OpenPifPaf[2] that is a bottom-up human pose estimation method, and compared the following two methods whose image selection mechanisms are different.

Random selection method

Randomly selects candidates from unlabeled image sequences and annotates them as training data. The uniform distribution is used for random selection process.

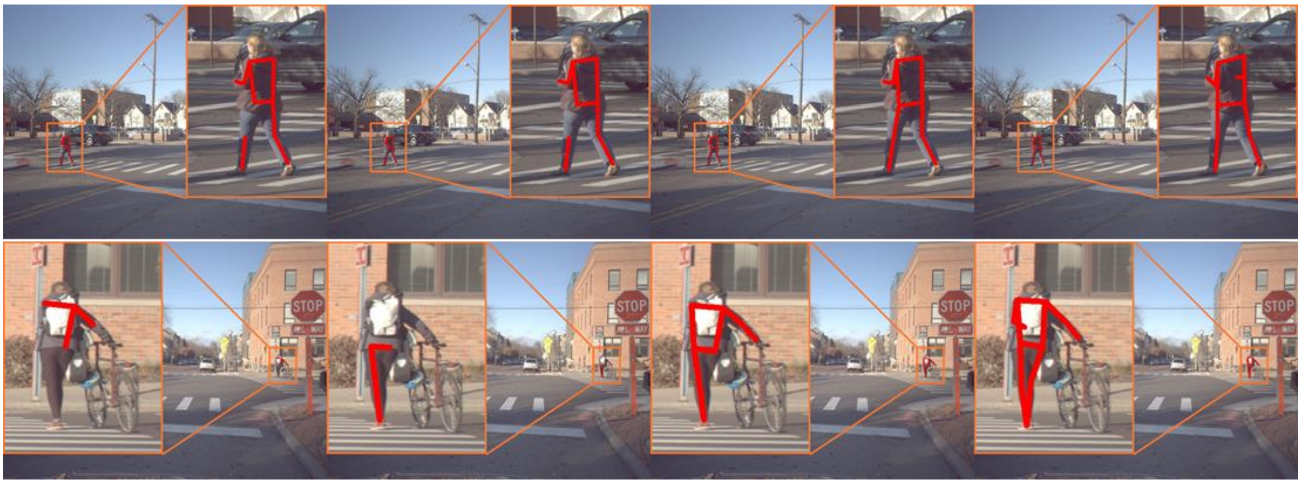
Proposed method

Automatically selects candidates from unlabeled image sequences in descending order of likelihood and annotates them as training data.

Average Precision (AP) was used as the evaluation metric, and the average of three experiments was calculated for evaluation. Object Keypoint Similarity (OKS)[9] was used to determine the correctness of the estimated pose for calculating the AP. OKS is an index that determines the correctness of the estimation based on the distance between the estimated and ground truths of the joints, the size of the person, and the weights determined for each joint.

Table 2. Changes of Average Precision (AP) [%] by annotating and adding unlabeled images.

Method	10% added	20% added	30% added
Random selection	48.1	52.3	52.2
Proposed	**49.0**	**52.6**	**53.2**



(a) Initial model (b) Random selection (10% added) (c) Proposed (10% added) (d) Ground Truth

Figure 3. Example of human pose estimation results.

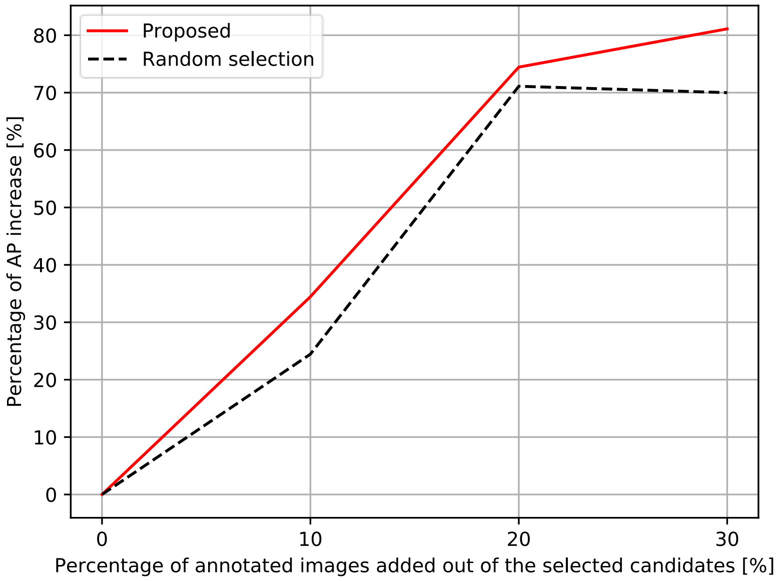


Figure 4. Percentage of AP increases for each method.

3.3 Results

Table 2 shows APs of each method when adding 10%, 20%, and 30% annotated images out of the candidates selected from unlabeled image sequences, and examples of estimated human poses are shown in Figure 3. The AP of the initial human pose estimation model was 45.9 (equivalent to 0% added), and the AP improved to 54.9 when adding all images (equivalent to 100% added). Figure 4 shows the increase of AP as a ratio to this 9.0 improvement in AP.

In the upper part of Figure 3, the right knee and the ankle were estimated as the joints of different persons by the random selection method, while they were correctly estimated as the joints of the same person by the proposed method. In the bottom row, we can see that the proposed method can estimate more joints than the random selection method. From Table **2**, we confirmed that the proposed method could improve the AP by considering the difference of the estimated joint positions between adjacent frames and the presence of the joints. This could be confirmed in all cases when 10%, 20%, and 30% of candidates were added in comparison with the random selection method. In particular, the highest improvement compared to the random selection method was observed when 30% of candidates were added, and the AP increased by 1.0.

From Figure 4, compared to the improvement of AP increase when all 3,188 unlabeled image sequences were annotated and added to the training data, the proposed method increased the AP by 34.4% when 319 (10%) images were added, by 74.4% when 638 (20%) images were added, and by 81.1% when 956 (30%) images were added. This indicates that the proposed method can efficiently improve the performance by adding a small number of training data.

4. CONCLUSION

In this paper, we proposed an active learning method that automatically selects candidates to be annotated manually for improving a human pose estimation method. The proposed method automatically selects images from unlabeled image sequences containing incorrectly estimated poses as additional annotation candidates based on the difference in joint positions of the estimated poses and the difference in the detected joints between adjacent frames. The experimental results showed that the proposed method is more efficient than a random selection method. Future work will include the development of a method that takes into account the differences in the intensity of human movement in an image sequence.

ACKNOWLEDGMENT

Parts of this research were supported by Grant-in-Aid for Scientific Research (Grant Number 17H00745).

REFERENCES

[1] Kreiss, S., Ramakrishna, V., Kanade, T. and Sheikh, Y., "PifPaf: Composite fields for human pose estimation," Proc. CVPR, 7424-4732 (2019).

[2] Kreiss, S., Bertoni, L., and Alahi, A., "OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association," arXiv Preprint arXiv:1805.06042 (2021).

[3] Wei, S.-E., Ramakrishna, V., Kanade, T. and Sheikh, Y., "Convolutional pose machines," Proc. CVPR, 4724-4732 (2016).

[4] Yan, S., Xiong, Y. and Lin, D., "Spatial temporal graph convolutional networks for skeleton-based action recognition," Proc. AAAI, 7444-7452 (2018).

[5] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y. and Hu, W., "Channel-wise topology refinement graph convolution for skeleton-based action recognition," Proc. ICCV, 13359-13368 (2021).

[6] Lewis, D. D. and Catlett, J., "Heterogeneous uncertainty sampling for supervised learning," Machine Learning Procs. 1994, 148-156 (1994).

[7] Liu, B. and Ferrari, V., "Active learning for human pose estimation," Proc. ICCV, 4363-4372 (2017).

[8] Xiu, Y., Li, J., Wang, H., Fang, Y. and Lu, C., "Pose flow: Efficient online pose tracking," Proc. BMVC, 53 (2018).

[9] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L., "Microsoft COCO: Common objects in context," Proc. ECCV, Part V, 740-755 (2014).

[10] Kim, W., Ramanagopal, M. S., Barto, C., Yu, M.-Y., Rosaen, K., Goumas, N., Vasudevan, R. and Johnson-Roberson, M., "PedX: Benchmark dataset for metric 3D pose estimation of pedestrians in complex urban intersections," IEEE Robot. Autom. Lett., 4(2), 1940-1947 (2019).