PAPER Special Section on Machine Vision and its Applications

Human Pose Annotation Using a Motion Capture System for Loose-Fitting Clothes

Takuya MATSUMOTO[†], **Kodai SHIMOSATO**[†], **Takahiro MAEDA**[†], **Tatsuya MURAKAMI**[†], **Koji MURAKOSO**[††], **Kazuhiko MINO**[††], ***Nonmembers***, ***and*** **Norimichi UKITA**[†a)], ***Senior Member***

SUMMARY This paper proposes a framework for automatically annotating the keypoints of a human body in images for learning 2D pose estimation models. Ground-truth annotations for supervised learning are difficult and cumbersome in most machine vision tasks. While considerable contributions in the community provide us a huge number of pose-annotated images, all of them mainly focus on people wearing common clothes, which are relatively easy to annotate the body keypoints. This paper, on the other hand, focuses on annotating people wearing loose-fitting clothes (e.g., Japanese Kimono) that occlude many body keypoints. In order to automatically and correctly annotate these people, we divert the 3D coordinates of the keypoints observed without loose-fitting clothes, which can be captured by a motion capture system (MoCap). These 3D keypoints are projected to an image where the body pose under loose-fitting clothes is similar to the one captured by the MoCap. Pose similarity between bodies with and without loose-fitting clothes is evaluated with 3D geometric configurations of MoCap markers that are visible even with loose-fitting clothes (e.g., markers on the head, wrists, and ankles). Experimental results validate the effectiveness of our proposed framework for human pose estimation.

key words: *human pose estimation, pose annotation, loose-fitting clothes*

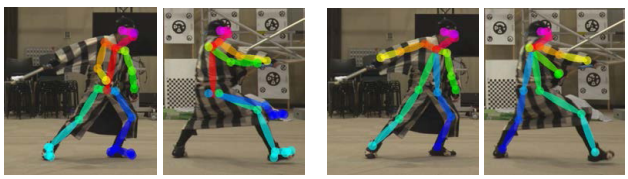
1. Introduction

Human pose estimation allows us to achieve a number of real-world applications such as CG human animation generation from videos and computer-supported study of physical skills such as sports and dances. While recent improvement of deep neural networks enables accurate pose estimation [1]–[5], they require a huge amount of supervised training data. The supervised data for human pose estimation is a set of images annotated with the keypoints of a human body (e.g., shoulders, wrists, knees, and ankles). This annotation is given manually to images (e.g., LSP [6] and MPII Human Pose [7]), in general, for 2D pose estimation where x–y image coordinates of each keypoint is estimated. Otherwise, incorrectly-annotated data are unavoidable in the automatic annotation (e.g., BBC Pose [8]) using pose estimation methods. For 3D pose estimation, the 3D coordinates of each keypoint can be measured by a Motion Capture system (MoCap) (e.g., HumanEva [9] and Human3.6M [10]), while it is difficult to use the MoCap in the wild.

Manuscript received August 23, 2019.
Manuscript revised January 28, 2020.
Manuscript publicized March 30, 2020.

[†]The authors are with Toyota Technological Institute, Nagoya-shi, 468–8511 Japan.
[††]The authors are with Zukun Lab, Toei Digital Center, Tokyo, 178–8666 Japan.
a) E-mail: ukita@toyota-ti.ac.jp

DOI: 10.1587/transinf.2019MVP0007



Pose estimated without additional annotations | Pose estimated with annotations on loose clothes

Fig. 1 Effects of additional pose annotations on loose-fitting clothes for human pose estimation. The pose estimation model obtained by our proposed annotation method correctly modifies erroneous poses of the original model.

However, it is difficult for the aforementioned manual and MoCap-based annotations to correctly localize the keypoints occluded by loose-fitting clothes such as Japanese Kimono and similar folk clothes. In this paper, we explore how to annotate such occluded keypoints under loose-fitting clothes for improving the pose estimation performance, as shown in Fig. 1, by utilizing a set of correctly-captured 3D keypoints. 3D keypoints can be captured by using the MoCap system in a standard manner where a person wearing tight-fitting clothes. Assume that similar body poses are observed both with tight-fitting and with loose-fitting clothes. Under this assumption, we project the keypoints captured with the tight-fitting clothes to images with the loose-fitting clothes.

Technical problems for the aforementioned annotation framework and our solutions are as follows:

Pose matching: We must match similar poses observed with tight- and loose-fitting clothes. Since this matching is difficult in images, as shown in "Image sequences" in Fig. 2, we employ the 3D coordinates of MoCap markers attached to visible endpoints (e.g., ankles) even with loose-fitting clothes. If the endpoints are localized in the same configuration in two different body poses, these poses may be similar to each other, as assumed in Inverse Kinematics.

Similar configurations of markers: Even if similar poses are captured both with tight- and loose-fitting clothes, these poses might be observed in different locations and orientations. Therefore, the geometric configurations of the markers are matched after they are spatially aligned.

The core contributions of this work are published in a conference [11]. Compared with this original version, this

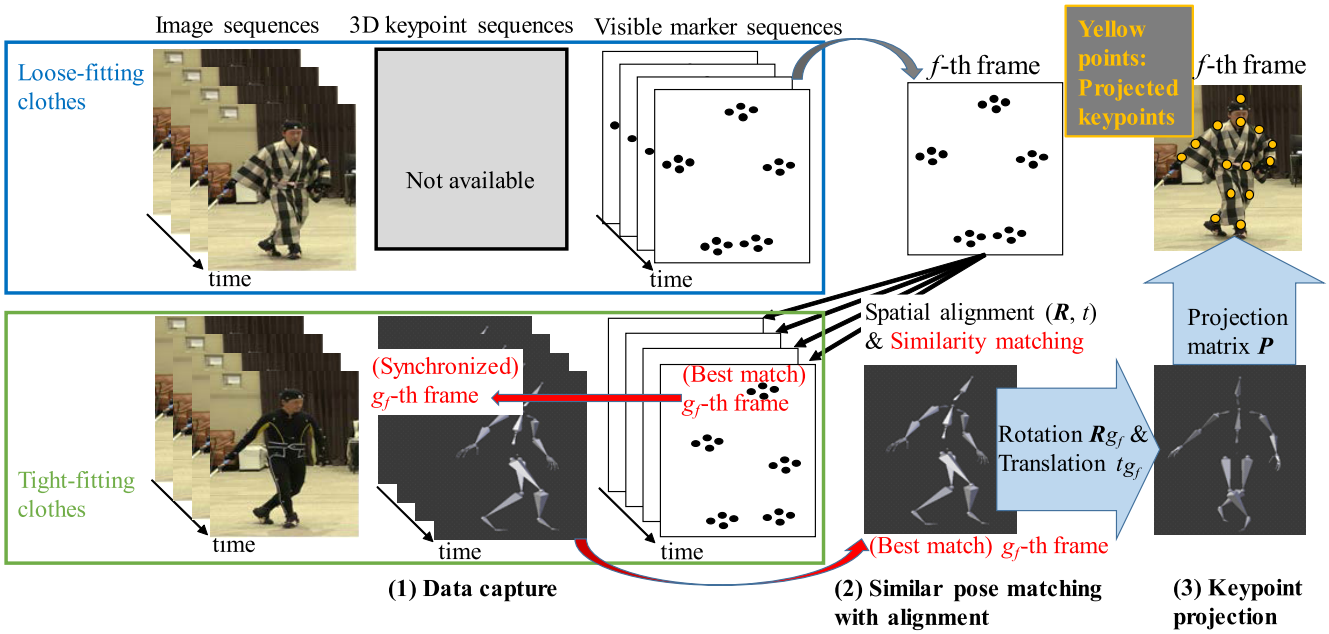


Fig. 2 Pipeline of the proposed method. In (1) data capture step, image sequences are captured in synchronization with a MoCap system. The MoCap system cannot measure the 3D keypoints of a human body in sequences with loose-fitting clothes, as shown by "Not available" in the figure. The MoCap system outputs visible markers as well as the 3D keypoints. (2) Pose matching with alignment finds that, for each frame in sequences with loose-fitting clothes (denoted by f-th frame), g_f-th frame in sequences with tight-fitting clothes is the most similar one in terms of the 3D configuration of the visible markers. Finally, the 3D keypoints in g_f-th frame are projected to f-th frame in (3) keypoint projection step.

paper presents more detailed results in comparative experiments.

2. Related Work

Around a decade ago, most 2D human pose estimation methods were based on pictorial structure models [12] and deformable part models [13]. While these models [12], [13] employ image-independent relationships between neighboring body parts for efficient computation, image-dependent relationships can improve the performance [14], [15]. Depth cameras including RGBD cameras allow us to estimate a human body pose more robustly and efficiently [16]. While depth images provide useful cues for pose estimation, their availability is much less than that of common RGB images.

The recent improvement of convolutional neural networks enables more accurate pose estimation even in RGB images [1]–[5]. As well as all machine learning based approaches, all of these pose estimation methods require huge training datasets (e.g., [7], [17]). Since erroneous pose annotations lead to failure in pose estimation, the annotations should be as correct as possible. While erroneous pose annotations can be modified during learning [18] in a similar manner to weakly-supervised learning, such approaches are insufficient for correct annotations. While human images annotated with correct keypoints can be synthesized by CG [19] and pose-annotated images can be differently textured so that a target person is dressed even with loose-fitting clothes in the image [19], it is known that the performance is limited if only such synthesized data is trained. Therefore, this paper proposes pose annotations on real images.

As mentioned in Introduction, pose annotation is difficult in particular for people wearing loose-fitting clothes. While recent deep neural networks for 2D keypoint detection [1], [2] and 3D body model matching [20] can predict a body pose even under loose-fitting clothes as shown in their experimental results, these methods do not explicitly cope with difficulty with loose-fitting-clothes. Therefore, their pose estimation accuracy for loose-fitting clothes is much lower than tight-fitting clothes. Pose annotation of such people in real images is addressed explicitly only by few previous methods. In [21], human body parts including loose-fitting clothes are automatically segmented based on colors painted on the clothes, which are difficult to be prepared. In addition, body keypoints cannot be correctly localized with the colored clothes. The keypoints under loose-fitting clothes are measured by a MoCap system using 3D gyroscopes, accelerometers, and magnetometers in [22]. However, the sensor drift error is unavoidable, and the magnetometers are also disturbed by metals around a subject.

3. Automatic Human Pose Annotation

Unlike previous approaches for human pose annotation introduced in Sect. 2, we propose to use a conventional optical MoCap system for capturing similar body poses with and

without loose-fitting clothes. Figure 2 shows the pipeline of our framework.

(1) Data capture (Section 3.1): For our proposed approach, similar poses must be included in training data with tight- and loose-fitting clothes. This assumption is easily guaranteed so that a subject is requested to behave as same as possible in these two different settings when training data are captured.
While the 3D coordinates of keypoints are measured by MoCap in the setting with tight-fitting clothes, they are not available in that with loose-fitting clothes. In our approach, however, optical markers are attached to the body also with loose-fitting clothes so that their locations are same as those with tight-fitting clothes. Some of these optical markers (e.g., those on the head, wrists, and ankles) are visible and fixed on the body even with loose-fitting clothes. As well as the aforementioned MoCap sequences, image sequences are captured simultaneously.

(2) Pose matching with alignment (Section 3.2): The 3D pose similar to the one observed with loose-fitting clothes is found from 3D poses with tight-fitting clothes captured by the MoCap system. Since this matching is difficult in an appearance domain, it is achieved by employing the 3D coordinates of optical markers that are visible even with loose-fitting clothes. We use the markers on the head, wrists, and ankles in our experiments under the assumption that these markers are visible in many frames. In order to match two 3D poses located in different positions and orientations in the MoCap coordinate system, these poses are spatially aligned.

(3) Keypoint projection (Section 3.3): Given the nearest neighbor 3D pose found from data with tight-fitting clothes. All keypoints at this frame are projected onto an image synchronized with the markers that are matched with this nearest neighbor pose. These projected keypoints are regarded as keypoint annotations.

3.1 Data Capture for Tight- and Loose-Fitting Clothes

In our data capture step, image and MoCap sequences are captured. While any cameras can be used for image capturing, we assume that the MoCap sequences are captured by an optical MoCap system.

A subject is requested to perform the same motions with tight- and loose-fitting clothes. While the 3D coordinates of body keypoints are measured in the setting with tight-fitting clothes, the keypoints are not available in that with loose-fitting clothes. However, for pose matching described in Sect. 3.2, optical markers are attached to the body also in the setting with loose-fitting clothes. In our experiments, the subject wears the loose-fitting clothes over the tight-fitting clothes with the markers.

If possible, it is better to synchronize cameras and a Mocap system. In our method, however, body keypoints captured with tight-fitting clothes are projected onto images with loose-fitting clothes, as described in Sect. 3.3. Since it is impossible to synchronize between the sequences of different observations, subtle time shifts between the image and the keypoints projected onto the image are unavoidable. It is also essentially impossible for the subject to repeat the completely same motions in the different observations. Therefore, hardware synchronization between cameras and MoCap is not necessarily required.

3.2 Pose Matching with Spatial Alignment

We have image and MoCap sequences with tight- and loose-fitting clothes. For each image observed with loose-fitting clothes, a 3D body pose captured in this image is matched with any 3D body pose captured with tight-fitting clothes.

Let N_m be the number of visible markers both in tight- and loose-fitting clothes, and $\boldsymbol{M}^{(l)}_{f,i} = \left(M^{(l)}_{f,i,x}, M^{(l)}_{f,i,y}, M^{(l)}_{f,i,z}, 1\right)^T$, where $i \in \{1, \cdots, N_m\}$, be the homogeneous coordinates of the i-th visible marker of a subject wearing loose-fitting clothes in f-th frame of a sequence. $\boldsymbol{M}^{(t)}_{g,i}$ denotes those with tight-fitting clothes. The markers attached to the same location (i.e., $\boldsymbol{M}^{(l)}_{f,i}$ and $\boldsymbol{M}^{(t)}_{g,i}$) are identified by the MoCap system.

In different observations, the location and orientation of the subject in the MoCap coordinate system may be changed. In order to spatially align two 3D body poses, the relative translation and rotation, denoted by $\boldsymbol{t}_{f,g}$ and $\boldsymbol{R}_{f,g}$ respectively, between f-th and g-th frames is computed based on the minimum mean square error as follows:

$$\boldsymbol{M}^{(l)}_{f} = \begin{bmatrix} \boldsymbol{R}_{f,g} & \boldsymbol{t}_{f,g} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \boldsymbol{M}^{(t)}_{g} \tag{1}$$

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{R}_{f,g} & \boldsymbol{t}_{f,g} \\ \boldsymbol{0}^T & 1 \end{bmatrix} = \boldsymbol{M}^{(l)}_{f} \boldsymbol{M}^{(t)^+}_{g} \tag{2}$$

$$\boldsymbol{M}^{(l)}_{f} = \left[\boldsymbol{M}^{(l)}_{f,1} \cdots \boldsymbol{M}^{(l)}_{f,N_m}\right]$$

$$\boldsymbol{M}^{(t)}_{g} = \left[\boldsymbol{M}^{(t)}_{g,1} \cdots \boldsymbol{M}^{(t)}_{g,N_m}\right]$$

Equation (2) is computed for each pair of $\boldsymbol{M}^{(l)}_{f}$ and $\boldsymbol{M}^{(t)}_{g}$. With $\boldsymbol{M}^{(t)'}_{g} = \boldsymbol{Q}\boldsymbol{M}^{(t)}_{g}$, dissimilarity between the spatially-aligned body poses is defined by the Mean Square Error (MSE) between all pairs of $\boldsymbol{M}^{(l)}_{f,i}$ and $\boldsymbol{M}^{(t)'}_{g,i}$:

$$E_{f,g} = \frac{1}{N_m} \sum_{i=1}^{N_m} \|\boldsymbol{M}^{(l)}_{f,i} - \boldsymbol{M}^{(t)'}_{g,i}\|^2 \tag{3}$$

If $\boldsymbol{M}^{(l)}_{f}$ and $\boldsymbol{M}^{(t)}_{g}$ come from different body poses, $\boldsymbol{t}_{f,g}$ and $\boldsymbol{R}_{f,g}$ are meaningless and the dissimilarity score, $E_{f,g}$ in (3), becomes larger. With this dissimilarity score, pose matching with spatial alignment is achieved as follows:

$$g_f = \arg\min_{g} E_{f,g} \tag{4}$$

where g_f denotes the frame in the tight-fitting clothes sequence that is most similar to f-th frame of the loose-fitting clothes sequence.

3.3 Keypoint Projection

All keypoints in g_f-th frame are measured by the MoCap system. These keypoints are projected onto f-th frame of the loose-fitting clothes sequence. A perspective projection matrix from the MoCap coordinate system to the 2D image coordinate system is computed with point correspondences between the 3D coordinates of MoCap markers and their 2D image coordinates [23]. The projected keypoints are utilized as human pose annotations for training human pose estimation models.

4. Experimental Results

(1) Data Capture

For capturing various kinds of free motions, all data was captured in a wide studio. Its dimension is 10m width × 7m depth × 2.5m height. We used a MoCap system consisting of 24 VICON T160 cameras (16 Megapixels). The resolution of RGB image sequences is 1920 × 1080 pixels.

Sample images with tight- and loose-fitting clothes are shown in Fig. 3, which are called tight- and loose-fitting datasets, respectively. In addition, we also prepared the Samurai film dataset, which was extracted from a real film.

Tight-fitting dataset: 3463 frames for training.

Loose-fitting dataset: 3744 and 1300 frames for training and test, respectively. The test frames come from sequences that are different from those used for the training frames in order to validate ability in model generalization.

Samurai film dataset: 174 test frames. A part of motions (200 images) in the tight- and loose-fitting training sets are imitations of motions in the Samurai film test set. Such training datasets make inference in the Samurai test set easy. On the other hand, the Samurai test set is more difficult than the loose-fitting test set due to noisy low-image quality, occlusion due to background objects (e.g., the rightmost samples in Fig. 6), and different subjects and clothes between training and test.

During our data capture step, one subject who performed a huge variety of motions was observed. Our datasets have the following challenging properties:

Variety: A huge variety of motions lead to difficulty in pose matching between tight- and loose-fitting datasets.

Asynchronicity: While the subject tried to behave as same as possible when with and without loose-fitting clothes, his location, orientation, and poses are different among sequences; It is impossible for the subject to completely spatially-align and temporally-synchronize the motions in different observations, in particular when the subject moves quickly, as shown in examples in Fig. 3: in both of the left and right examples, head and body orientations are different between images with tight- and loose-fitting clothes.

Fig. 3 Pose-matched frames. Each pair shows the best matching results obtained by the proposed pose matching method. Keypoint projected onto each image are indicated by red points. It can be seen that these matched images are similar enough to use the projected keypoints for annotations.

On the other hand, our datasets are insufficient for validating generalizability of the proposed method because (1) only one subject is observed for loose- and tight-fitting datasets and (2) the subject imitated motions observed in the Samurai film test set. While most machine-learning algorithms are designed to improve generalizability from the dataset, the focus of this work is not such generalizability but exploration of possibility of automatically annotating a person wearing loose-fitting clothes. In addition to the issue of generalizability, the subject whose motions were observed for the tight- and loose-fitting datasets is a professional actor. While he is good at motion imitation, such an actor is not available for general purposes. Despite such limited applicability, our proposed method is useful as is for several professional uses such as 2D film analysis, which is shown with the Samurai film dataset, for producing 3D CG films.

With our proposed method, all training images were automatically pose-annotated. Only for evaluation, all images including training and test images were manually annotated. The manual annotations in the training and test images are used for evaluating the effect of our spatial alignment scheme and for evaluating the performance on pose estimation, respectively.

(2) Pose Matching with Spatial Alignment and Keypoint Projection

For each frame with loose-fitting clothes, its best match frame with tight-fitting clothes is found. For this pose matching, 23 markers were used in total; five, eight, and ten optical markers are attached to the visible regions of the head, wrists, and ankles, respectively, in the MoCap system. Figure 3 shows the examples of this pose matching.

Based on this pose matching, a set of 3D keypoints captured with tight-fitting clothes in each frame was projected to its corresponding frame with loose-fitting clothes. The mean distance between the projected keypoints and their corresponding ground-truth positions is shown in Table 1. For comparison, the mean distance obtained without our spatial alignment is also shown. Table 1 validates the effectiveness of the spatial alignment; 26.7 pixel error without spatial alignment vs 19.6 pixel error with spatial alignment in total.

Table 1 The distance (pixels) between the automatically-annotated keypoint and its ground-truth (denoted by d_w). Its mean over all keypoints and all frames is shown. For validating the effect of the spatial alignment, defined by $\boldsymbol{Q}$ in (2), for pose matching, the distance in case of no spatial alignment (denoted by d_o) is also shown. The bottom row, "error reduction rate", is computed to be $\frac{d_w}{d_o} \times 100$.

	head	shoulders	elbows	wrists	hips	knees	ankles	total
without spatial alignment	26.7	24.3	32.0	22.9	30.9	26.1	24.0	26.7
with spatial alignment	19.6	15.5	21.8	14.5	26.6	21.3	18.3	19.6
error reduction rate	27%	36%	32%	37%	14%	18%	24%	27%

Table 2 PCKh-0.5 evaluation [7] on our loose-fitting dataset. The best score obtained on each dataset in each column is colored by red.

	head	shoulders	elbows	wrists	hips	knees	ankles	total
Baseline ([2] without finetune)	100	99.0	67.2	48.5	96.3	87.3	85.9	82.2
Baseline ([2] finetuned with tight clothes)	100	98.3	83.5	80.9	97.8	95.9	98.1	93.0
Ours ([2] finetuned with loose clothes)	100	99.5	93.2	85.0	99.3	96.9	98.0	95.5
Baseline ([24] without finetune)	93.5	97.8	83.7	69.4	93.5	84.7	95.1	87.8
Baseline ([24] finetuned with tight clothes)	82.5	96.2	79.9	36.9	92.3	86.3	90.3	80.5
Ours ([24] finetuned with loose clothes)	100	98.8	92.6	91.1	98.5	96.7	97.0	96.1

Table 3 PCKh-0.5 evaluation [7] on the Samurai film dataset. The best score obtained on each dataset in each column is colored by red.

	head	shoulders	elbows	wrists	hips	knees	ankles	total
Baseline ([2] without finetune)	64.9	50.9	12.9	9.5	25.9	7.5	1.7	21.7
Baseline ([2] finetuned with tight clothes)	48.3	53.7	20.4	25.9	46.3	24.7	6.3	31.0
Ours ([2] finetuned with loose clothes)	68.4	63.8	42.8	29.0	47.7	27.6	13.8	39.8
Baseline ([24] without finetune)	48.3	81.0	34.5	37.4	62.1	56.6	42.2	52.0
Baseline ([24] finetuned with tight clothes)	85.1	77.9	48.0	48.6	55.5	58.3	37.4	56.6
Ours ([24] finetuned with loose clothes)	83.9	78.7	48.6	50.3	68.4	48.6	38.2	57.6

Table 4 PCKh-0.5 evaluation [7] on the Samurai film dataset. For our proposed finetuning, a part of loose-fitting dataset were used so that only motions imitating the Samurai film set were trained.

	head	shoulders	elbows	wrists	hips	knees	ankles	total
Ours ([2] finetuned with imitation motions)	47.1	44.0	0.0	0.0	44.0	14.7	0.0	19.4

The set of the keypoints in the images is employed as the additional human pose annotations for finetuning the pose estimation model. In what follows, this pose estimation model is evaluated in order to validate how effective the matching results are for improving human pose estimation.

(3) Pose Estimation

Pose estimation methods proposed in [2] and [24] are used for evaluation. Their pose estimation models were pretrained with the COCO dataset [17] and the VGG-19. The pretrained models were given by the authors of [2] and [24]. For our proposed method, these models were finetuned by our training images with loose-fitting clothes. Additional experiments with finetuning by our training images with tight-fitting clothes were also conducted. The parameters used in this finetuning are as follows:

Cao et al. [2]: SGD with learning rate = $4.0e^{-5}$, momentum = 0.9, weight decay = $5.0e^{-4}$, batch-size = 10, and epochs = 1.

Xiao et al. [24]: Adam with learning rate = $1.0e^{-3}$, momentum = 0.9, weight decay = $1.0e^{-4}$, batch-size = 32, and epochs = 3.

Figure 4 shows the visualized results of the finetuned model using [2] on the loose-fitting dataset. For comparison, the results of the baseline (i.e., the pretrained model using [2]) are also shown. The quantitative results evaluated by PCKh-0.5 [7] are shown in Table 2 †. The PCKh curves with [2] are also shown in Fig. 5. It can be seen that our proposed model outperforms the original model and the model finetuned by tight-fitting clothes in most PCKh thresholds in all keypoints. In particular, the detection rates in the elbows and wrists, which are difficult to be detected by the original model, are improved significantly.

In order to validate the generalizability of the model

†In our experiments, the head consists of "neck". While "neck" is not annotated in the COCO dataset, the mean of two shoulders is regarded as its position in accordance with [2].

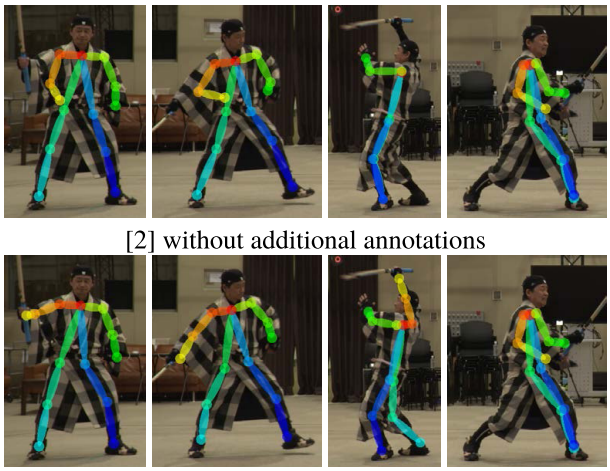


Fig. 4 Visualization of poses estimated by [2] on our loose-fitting dataset.

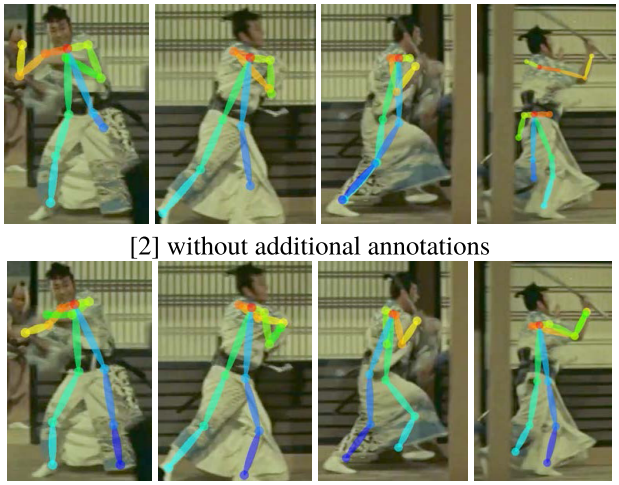


Fig. 6 Visualization of poses estimated by [2] on our Samurai film dataset.

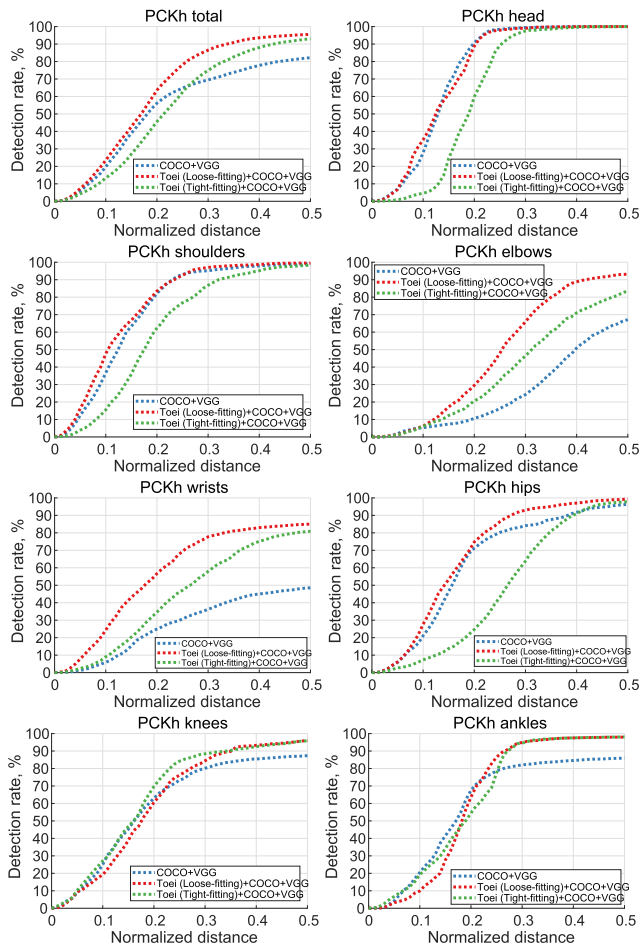


Fig. 5 PCKh curves of [2] with the original model and our model fine-tuned by our loose-fitting dataset. The curves of all keypoints and their mean (i.e., total in Tables 2 and 3) are shown.

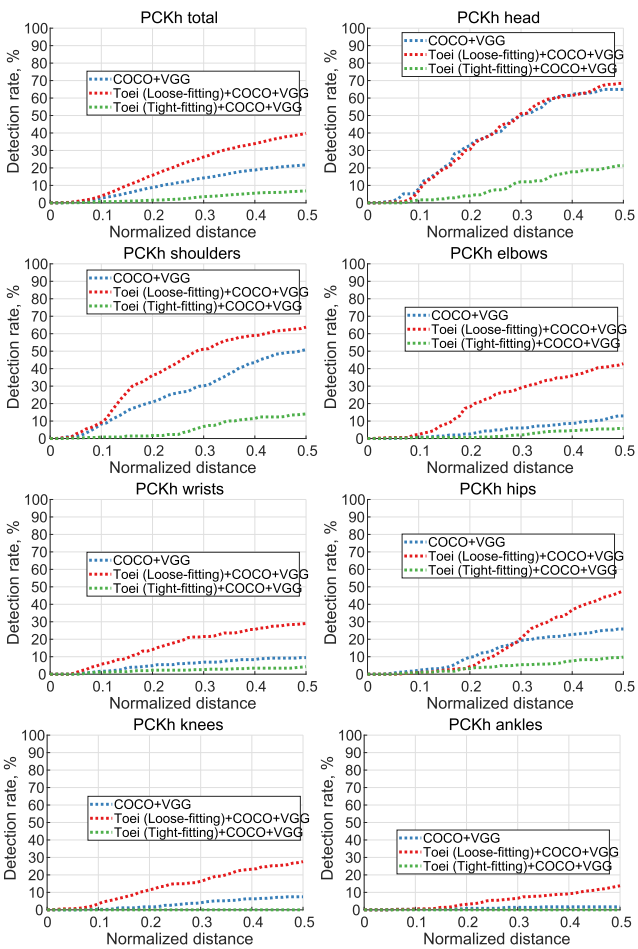


Fig. 7 PCKh curves of [2] with the original model and our model fine-tuned by additional pose annotations on the Samurai film dataset. The curves of all keypoints and their mean are shown.

finetuned with our loose-fitting dataset, we applied these models to real Samurai film sequences (Fig. 6). Table 3, Fig. 6, and Fig. 7 show the results of PCKh-0.5 evaluation, visualized results, and PCKh curves, respectively. Pose estimation on the Samurai film dataset is tough because occlusion with long sleeves and hem is more severe than in the loose-fitting dataset. In addition, the image quality of the Samurai film dataset is much lower than the loose-fitting dataset we captured. Due to these difficult problems, the detection rates of the original model in the Samurai film dataset are quite lower in all keypoints than in the loose-fitting dataset. Training images with tight-fitting clothes cannot support pose estimation in the Samurai film dataset. However, our proposed model using training images with loose-fitting clothes can improve both of the quantitative and qualitative results.

Table 4 also shows the result of our proposed method with [2] finetuned by a part of the loose-fitting dataset. For this finetuning, training images are selected so that only motions imitating those observed in the Samurai film test set are trained. We call this finetuned model M_i. This experiment was conducted for investigating whether or not our proposed method performs well due to overfitting to the test set. It can be seen that M_i is inferior to our proposed method finetuned by all training images, which is shown in the third row of Table 3. On the contrary, M_i is inferior to the pretrained model, which is shown in the first row of Table 3. Recall that (1) the number of these imitation images used for M_i is 200 and (2) loose-fitting clothes and motions observed in the loose-fitting training set are just the imitations of the Samurai film dataset. We interpret this result as meaning that a small number of training images cannot improve the test performance even if these training images are similar to the test images.

5. Concluding Remarks

This paper proposed a framework for automatic pose annotation of people wearing loose-fitting clothes. In order to annotate the body keypoints under the loose-fitting clothes in images, we project the 3D coordinates of the keypoints without loose-fitting clothes captured by a MoCap system. Pose similarity between bodies with and without loose-fitting clothes in different observations is achieved based on 3D geometric configurations of visible MoCap markers.

Future work includes using temporal cues for pose matching. The effectiveness of the temporal cues is validated (e.g., using latent models [22], [25]–[28] and using deep networks [29], [30]) and is expected to be useful in our proposed method also. While the proposed method just projects the best matched pose to an image, the projected pose can be further validated by keypoint connectivity in the appearance domain [2], [15]. For modeling more various appearances of loose-fitting clothes, semi/weakly-supervised learning is also important for [31]. Extension to 3D pose estimation [32]–[34] is also an important research direction.

For our experiments, a professional actor was requested to behave as same as possible with tight- and loose-fitting clothes for training. In addition, for estimating poses observed in the Samurai film set, the actor imitated motions observed in this set. For more general scenarios where any person can move freely both for training and test data, a more huge variety of body poses must be captured for training. Generalizability of a pose estimation method is also required for coping with variation between training and test data. These topics are also included in future work.

This work was partly supported by JSPS KAKENHI Grant Number 19K12129.

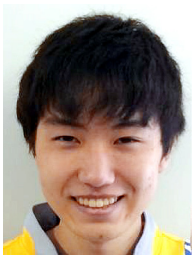
References

[1] S.E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," CVPR, 2016.
[2] Z. Cao, T. Simon, S.E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," CVPR, 2017.
[3] Y. Kawana, N. Ukita, J. Huang, and M. Yang, "Ensemble convolutional neural networks for pose estimation," CVIU, vol.169, pp.62–74, April 2018.
[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," CVPR, 2019.
[5] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated multi-person tracking in the wild," CVPR, 2017.
[6] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," BMVC, 2010.
[7] M. Andriluka, L. Pishchulin, P.V. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," CVPR, 2014.
[8] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," IJCV, vol.110, no.1, pp.70–90, 2014.
[9] L. Sigal, A.O. Balan, and M.J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," IJCV, vol.87, no.1-2, pp.4–27, 2010.
[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," PAMI, vol.36, no.7, pp.1325–1339, July 2014.
[11] T. Matsumoto, K. Shimosato, T. Maeda, T. Murakami, K. Murakoso, K. Mino, and N. Ukita, "Automatic human pose annotation for loose-fitting clothes," MVA, 2019.
[12] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial structures for object recognition," IJCV, vol.61, no.1, pp.55–79, 2005.
[13] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," PAMI, vol.32, no.9, pp.1627–1645, Sept. 2010.
[14] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," ECCV, 2010.
[15] N. Ukita, "Articulated pose estimation with parts connectivity using discriminative local oriented contours," CVPR, 2012.
[16] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," CVPR, 2011.
[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft COCO: common objects in context," ECCV, 2014.
[18] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," CVPR, 2011.
[19] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved CNN supervision," 3DV, 2017.

[20] R.A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," CVPR, 2018.
[21] N. Ukita, R. Tsuji, and M. Kidode, "Real-time shape analysis of a human body in clothing using time-series part-labeled volumes," ECCV, 2008.
[22] N. Ukita, M. Hirai, and M. Kidode, "Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints," ICCV, 2009.
[23] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2 ed., Cambridge University Press, New York, NY, USA, 2003.
[24] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," ECCV, 2018.
[25] J.M. Wang, D.J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," PAMI, vol.30, no.2, pp.283–298, Feb. 2008.
[26] N. Ukita and T. Kanade, "Gaussian process motion graph models for smooth transitions among multiple actions," CVIU, vol.116, no.4, pp.500–509, 2012.
[27] N. Ukita, "Simultaneous particle tracking in multi-action motion models with synthesized paths," Image Vision Comput., vol.31, no.6-7, pp.448–459, 2013.
[28] K. Morimoto, Y. Matsuyama, and N. Ukita, "Continuous action recognition by action-specific motion models," MVA, pp.323–326, 2013.
[29] D. Zhang, G. Guo, D. Huang, and J. Han, "Poseflow: A deep motion representation for understanding human behaviors in videos," CVPR, 2018.
[30] H. Coskun, D.J. Tan, S. Conjeti, N. Navab, and F. Tombari, "Human motion analysis with deep metric learning," ECCV, 2018.
[31] N. Ukita and Y. Uematsu, "Semi- and weakly-supervised human pose estimation," CVIU, vol.170, pp.67–78, May 2018.
[32] M.R.I. Hossain and J.J. Little, "Exploiting temporal information for 3d human pose estimation," ECCV, 2018.
[33] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3d human pose from structure and motion," ECCV, 2018.
[34] C. Li and G.H. Lee, "Generating multiple hypotheses for 3d human pose estimation with mixture density network," CVPR, 2019.

Takuya Matsumoto is a bachelor course student of Toyota Technological Institute, Japan. His research theme is human pose estimation and 3D object recognition.

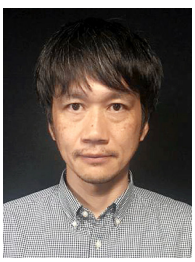
Kodai Shimosato is a bachelor course student of Toyota Technological Institute, Japan. His research theme is human pose estimation and remote sensing.



Takahiro Maeda is a master course student of Toyota Technological Institute, Japan. His research theme is human motion animation.

Tatsuya Murakami was a bachelor course student of Toyota Technological Institute, Japan. His research theme was human pose estimation.

Koji Murakoso is a Producer at Zukun Laboratory, TOEI CO.,LTD. He is mainly engaged in digital human research and management of various research.



Kazuhiko Mino is a chief director at Zukun Laboratory, TOEI CO.,LTD. Main fields are CG, VFX, XR, and especially, digital human and virtual production.

Norimichi Ukita is a professor at the graduate school of engineering, Toyota Technological Institute, Japan (TTI-J). He received the Ph.D. degree in Informatics from Kyoto University, Japan, in 2001. After working for five years as an assistant professor at NAIST, he became an associate professor in 2007 and moved to TTI-J in 2016. He was a research scientist of PRESTO, JST during 2002–2006. He was a visiting research scientist at Carnegie Mellon University during 2007–2009. His main research interests are multi-object tracking and human pose and activity estimation.