16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Automatic Human Pose Annotation for Loose-fitting Clothes

Takuya Matsumoto† Kodai Shimosato† Takahiro Maeda† Tatsuya Murakami†
Koji Murakoso‡ Kazuhiko Mino‡ Norimichi Ukita†
†Toyota Technological Institute ‡Zukun Lab, Toei Digital Center
{sd16080,sd16041,sd15082,ukita}@toyota-ti.ac.jp

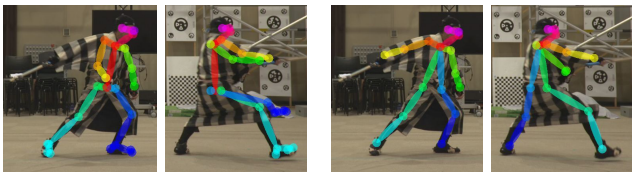
Abstract

This paper proposes a framework for automatically annotating the keypoints of a human body in images for learning 2D pose estimation models. 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 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. Experimental results validate the effectiveness of our proposed framework for human pose estimation.

1 Introduction

Human pose estimation allows us to achieve a number of real-world applications such as image/video retrieval. While recent improvement of deep neural networks enables accurate pose estimation [2, 9], 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 [7] and MPII Human Pose [1]), 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 [3]) 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 [13] and Human3.6M [6]), while it is difficult to use the MoCap in the wild.

However, it is difficult for the aforementioned manual and MoCap-based annotations to correctly localize the keypoints occluded by loose-fitting clothes. In this paper, we explore how to annotate such occluded keypoints for improving the pose estimation performance, as shown in Figure 1. 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.



Pose estimated without additional annotations

Pose estimated with annotations on loose clothes

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

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

Pose matching: We must match similar poses observed with tight-fitting and loose-fitting clothes. Since this matching is difficult in images, as shown in "Image sequences" in Figure 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-fitting 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.

2 Related Work

The recent improvement of convolutional neural networks enables accurate pose estimation even in RGB images [2, 9]. All of these pose estimation methods

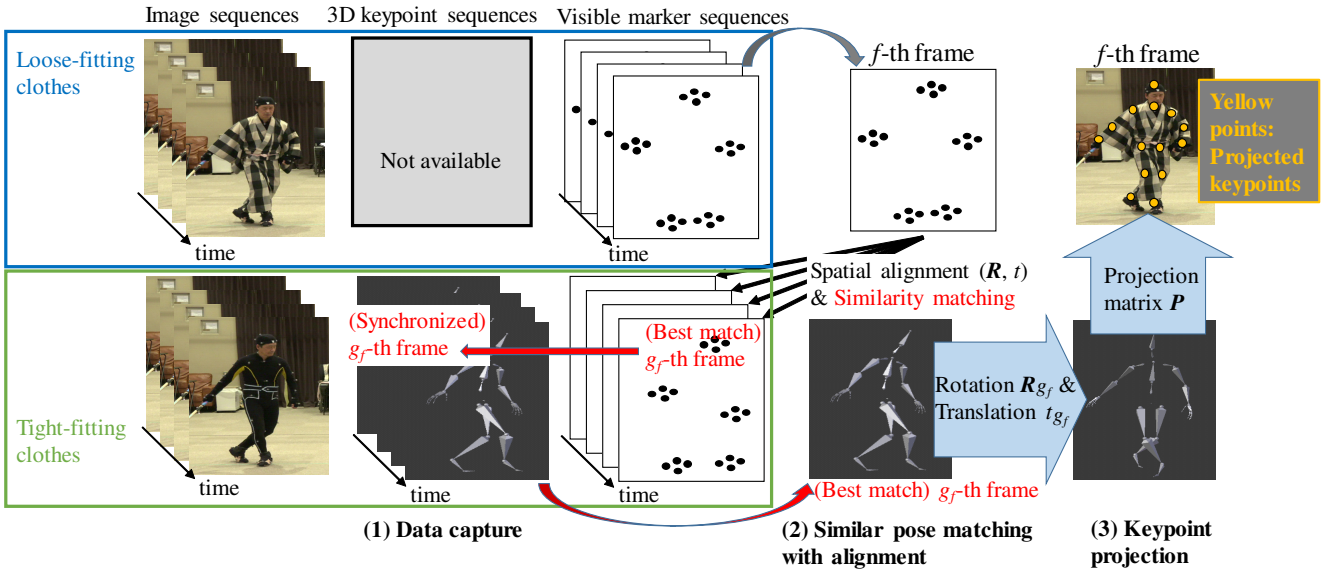


Figure 2. Pipeline of the proposed framework. In (1) data capture step, image sequences are captured in synchronization with the MoCap system. The MoCap system cannot measure the 3D keypoints of a human body in sequences with loose-fitting clothes, as shown by "Not available" in the figure. The MoCap 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.

require huge training datasets (e.g., [1, 10]). 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 [8] 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 [11], it is known that the performance is limited if only such synthesized data is trained. Therefore, this paper proposes pose annotations on real images.

Pose annotation is difficult in particular for people wearing loose-fitting clothes. Pose annotation of such people in real images (not in CGs) is addressed explicitly by few previous methods. In [17], human body parts including loose-fitting clothes are automatically segmented based on colors painted on the clothes, which are difficult to be prepared. The keypoints under loose-fitting clothes are measured by a MoCap system using 3D gyroscopes, accelerometers, and magnetometers in [15]. However, the sensor drift error is unavoidable, and the magnetometers are also disturbed by metals around a subject.

3 Automatic Human Pose Annotation for Supervised Learning

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-fitting 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 As well as MoCap sequences including 3D keypoint and visible marker 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. This matching is achieved with the 3D coordinates of visible optical markers. 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 in different positions and orientations, 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 on this images.

3.1 Data Capture for Tight-fitting 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-fitting 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 Section 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 fully synchronize cameras and a Mocap system. However, for automatic human pose annotation, body keypoints captured with tight-fitting clothes are projected onto images with loose-fitting clothes, as described in Section 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 the cameras and the MoCap system is not necessarily required.

3.2 Pose Matching with Spatial Alignment

We have image and MoCap sequences with tight-fitting 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-fitting and loose-fitting clothes, and $\boldsymbol{M}_{f,i}^{(l)} = \left(M_{f,i,x}^{(l)}, M_{f,i,y}^{(l)}, M_{f,i,z}^{(l)}, 1\right)^T$, 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. $\boldsymbol{M}_{g,i}^{(t)}$ denotes those with tight-fitting clothes. The markers attached to the same location (i.e., $\boldsymbol{M}_{f,i}^{(l)}$ and $\boldsymbol{M}_{g,i}^{(t)}$) 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:

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

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

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

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

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

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

If $\boldsymbol{M}_f^{(l)}$ and $\boldsymbol{M}_g^{(t)}$ 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:

$$g_f = \arg\min_g E_{f,g} \quad (4)$$

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 [5]. The projected keypoints are utilized as human pose annotations for training human pose estimation models.

4 Experimental Results

Our proposed method is evaluated with the performance on human pose estimation using with and without automatically-annotated keypoints in sequences with loose-fitting clothes.

Data Capture For capturing various kinds of free motions, all data was captured in a wide studio (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.

Table 1. The distance (pixels) between the automatically-annotated keypoint and its ground-truth (denoted by d_w) is computed, and 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%

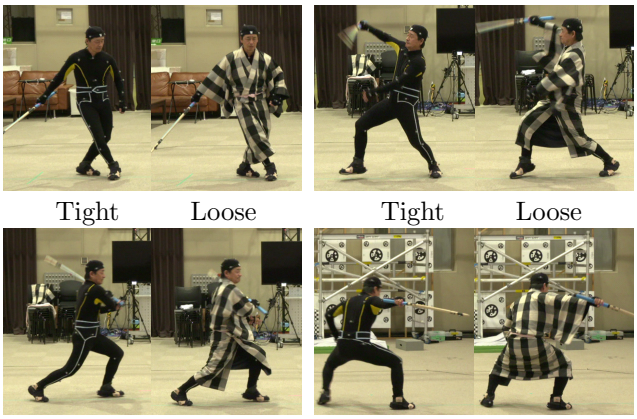


Figure 3. Pose-matched frames. Each pair shows the best matching results obtained by the proposed pose matching method.

Sample images with tight- and loose-fitting clothes are shown in Figure 3, which are called tight- and loose-fitting datasets, respectively. In addition to these two datasets, 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 training frames come from sequences that are different from those used for the test frames in order to validate ability in model generalization.

Samurai film dataset: 174 frames for test.

During our data capture step, one subject who performed a huge variety of motions was observed. This dataset has the following challenging properties:

Variety: A huge variety of motions lead to large differences between training and test data, which result in difficulty in pose estimation.

Complexity: The subject imitated the complex motions of Samurai film actors in order to evaluate our proposed method on real films as well as on our loose-fitting test dataset.

Asynchronicity: 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 the upper-right example in Figure 3.

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.

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 several 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 projected keypoints are also shown in Figure 3. 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.

Pose Estimation Pose estimation methods proposed in [2] and [20] are used for evaluation. Their pose estimation models were pretrained with the COCO dataset [10] and the VGG-19 model. The pretrained models were given by the authors of [2] and [20]. For our experiments, these models were finetuned by our training images with loose-fitting clothes. The parameters used in this finetuning are as follows:

[2]: SGD with learning rate = $4.0e^{-5}$, momentum = 0.9, and weight decay = $5.0e^{-4}$.

[20]: Adam with learning rate = $1.0e^{-3}$, momentum = 0.9, and weight decay = $1.0e^{-4}$.

Table 2. PCKh-0.5 evaluation [1] 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
Proposed method (with finetune)	100	99.5	93.2	85.0	99.3	96.9	98.0	95.5
Baseline [20] (without finetune)	93.5	97.8	83.7	69.4	93.5	84.7	95.1	87.8
Proposed method (with finetune)	100	98.8	92.6	91.1	98.5	96.7	97.0	96.1

Table 3. PCKh-0.5 evaluation [1] 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
Proposed method (with finetune)	68.4	63.8	42.8	29.0	47.7	27.6	13.8	39.8
Baseline [20] (without finetune)	48.3	81.0	34.5	37.4	62.1	56.6	42.2	52.0
Proposed method (with finetune)	83.9	78.7	48.6	50.3	68.4	48.6	38.2	57.6

Figure 4 shows the visualized results of the finetuned pose estimation 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 [1] are shown in Table 2 [1]. The PCKh curves with [2] are also shown in Figure 5. It can be seen that our proposed model outperforms the original model in all PCKh thresholds.

In order to validate the generalized performance of the pose estimation model finetuned with our loose-fitting dataset, we applied these models to real Samurai film sequences (Figure 6). Table 3, Figure 6, and Figure 7 show the results of PCKh-0.5 evaluation, visualized pose estimation results, and PCKh curves, respectively. Although (1) pose estimation on this dataset is tough due to severe occlusion with long sleeves and hem and (2) probably finetuning using our loose-fitting dataset is not generalized for other clothes and motions sufficiently, it can be seen that our proposed model is, in total, improved compared with the original model in both of the quantitative and qualitative results.

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.

Future work includes using temporal cues for pose matching because the proposed method achieves only framewise matching. The effectiveness of the temporal cues is validated (e.g., using latent models [19, 15, 16, 12] and using deep networks [21, 4]) 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 [14, 2]. In order to learn more data for modeling various appearances of loose-fitting clothes, semi/weakly-supervised learning is also important for [18].

This work was supported by the incubation program of Kyoto University.

[1] 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].

References

[1] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 2, 5

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 4, 5, 6

[3] J. Charles, T. Pfister, M. Everingham, and A. Zisserman. Automatic and efficient human pose estimation for sign language videos. *IJCV*, 110(1):70–90, 2014. 1

[4] H. Coskun, D. J. Tan, S. Conjeti, N. Navab, and F. Tombari. Human motion analysis with deep metric learning. In *ECCV*, 2018. 5

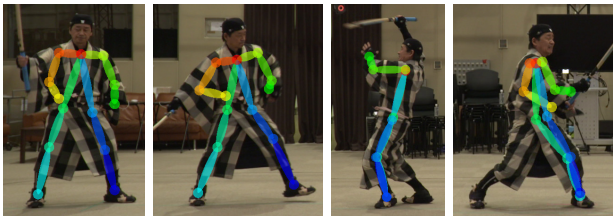
[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003. 3

[6] 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*, 36(7):1325–1339, 2014. 1

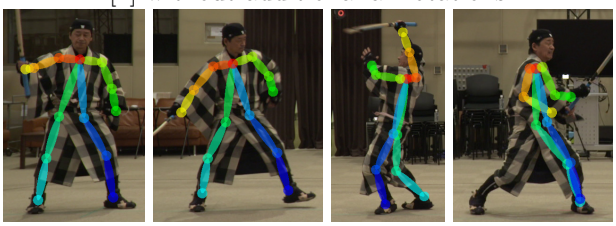
[7] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 1

[8] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 2

[9] Y. Kawana, N. Ukita, J. Huang, and M. Yang. Ensemble convolutional neural networks for pose estimation. *CVIU*, 169:62–74, 2018. 1



[2] without additional annotations



[2] with additional annotations on loose clothes

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

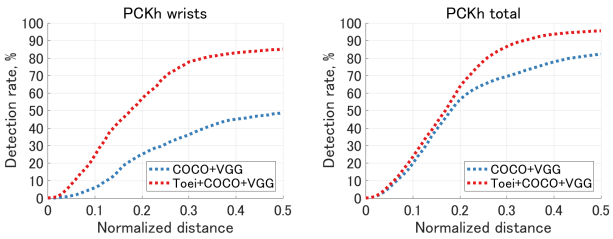
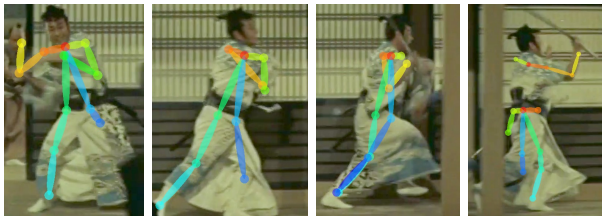
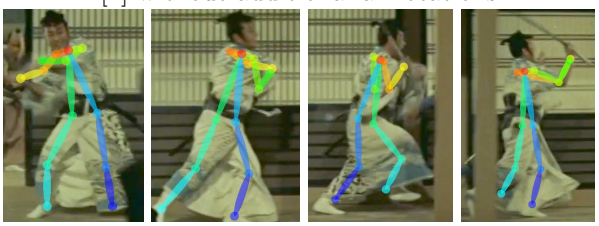


Figure 5. PCKh curves of [2] with the original model and our model finetuned by our loose-fitting dataset. The curves of the mean of all keypoints (i.e., total in Tables 2 and 3) and the wrists, which are most difficult to be localized, are shown.



[2] without additional annotations



[2] with additional annotations on loose clothes

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

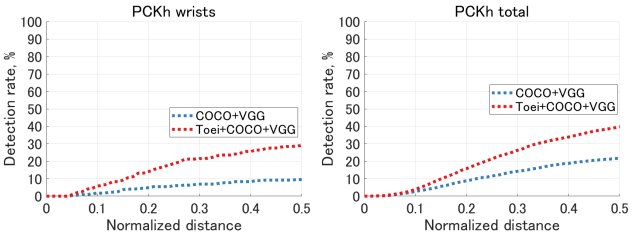


Figure 7. PCKh curves of [2] with the original model and our model finetuned by additional pose annotations on the Samurai film dataset.

[10] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 2, 4

[11] 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. In *3DV*, 2017. 2

[12] K. Morimoto, Y. Matsuyama, and N. Ukita. Continuous action recognition by action-specific motion models. In *MVA*, 2013. 5

[13] 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*, 87(1-2):4–27, 2010. 1

[14] N. Ukita. Articulated pose estimation with parts connectivity using discriminative local oriented contours. In *CVPR*, 2012. 5

[15] N. Ukita, M. Hirai, and M. Kidode. Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints. In *ICCV*, 2009. 2, 5

[16] N. Ukita and T. Kanade. Gaussian process motion graph models for smooth transitions among multiple actions. *CVIU*, 116(4):500–509, 2012. 5

[17] N. Ukita, R. Tsuji, and M. Kidode. Real-time shape analysis of a human body in clothing using time-series part-labeled volumes. In *ECCV*, 2008. 2

[18] N. Ukita and Y. Uematsu. Semi- and weakly-supervised human pose estimation. *CVIU*, 170:67–78, 2018. 5

[19] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *PAMI*, 30(2):283–298, 2008. 5

[20] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 4, 5

[21] D. Zhang, G. Guo, D. Huang, and J. Han. Poseflow: A deep motion representation for understanding human behaviors in videos. In *CVPR*, 2018. 5