04-25

15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Domain Adaptation of Articulated Pose Estimation via Synthetic Pose Prior

Kazuhiko Murasaki, Haruka Yonemoto, Kyoko Sudo and Tetsuya Kinebuchi
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847 Japan
{ murasaki.kazuhiko, yonemoto.haruka, sudo.kyoko, kinebuchi.t }@lab.ntt.co.jp

Abstract

This paper proposes an articulated pose estimation method based on the pose prior for adaptation that is scene specific. In this research field, various approaches to estimate articulated human pose have been proposed and many researchers have tried to improve pose estimation accuracy for shared datasets. On the other hand, it is not common to use datasets without any labeling for realizing domain adaptation to different scenes even though it is urgently needed. Domain adaptation without labeled data is the key problem because when the user wants to estimate human pose captured in a specific scene, it is too costly to make a labeled dataset for training the estimator. We tackle this problem by proposing the novel approach of using synthetic pose prior, which is made by simulating probable poses in the target scene. We consider the likelihood of human pose distribution generated from various motion capture data sets and environmental knowledge about the scene in addition to training a basic appearance-based pose estimator by using a labeled dataset. Using the likelihood of joint positions specific to the scene, more appropriate results are yielded. In experiments, we adapt an estimator trained by images captured from various perspectives to suit the target images captured from a fixed perspective. As the proposed method takes advantage of the bias caused by fixed perspective, it offers improved estimation accuracy.

1 Introduction

Articulated human pose estimation has been intently studied over the last decade in the computer vision field, but many problems remain. Especially recently, deep learning has significantly improved pose estimation accuracy, and a lot of novel approaches using deep learning have been proposed [1, 2]. There are also various open datasets for fair evaluation of human pose estimation, not only video captured in the laboratory [3], but also photos of athletes [4], screen captures of movies [5] and so on. Although many researchers use these open datasets to assess estimation accuracy, few papers have addressed domain adaptation for different scenes due to the difficulty of finding labeled datasets for training. Some papers have reported the estimation accuracy possible with adapting domain-different data [2], but they did not propose any specific method for domain adaptation. However, domain adaptation to specific scenes without labeled data is a problem of practical importance because when the user tries to estimate human pose captured in a specific scene, it is too costly to make a labeled dataset for training the estimator.

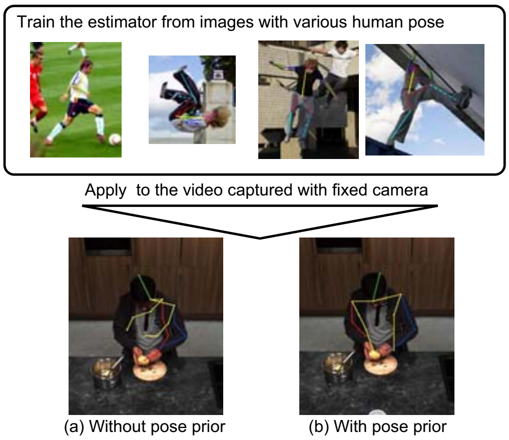
Train the estimator from images with various human pose
Apply to the video captured with fixed camera
(a) Without pose prior
(b) With pose prior

Figure 1. Domain adaptation of articulated pose estimation model trained with LSP dataset to Cooking Activities dataset. (a) Chen and Yuille's method without any prior. (b) Our method considering pose prior of the specific view.

In this paper, we propose a novel domain adaptation approach for articulated pose estimation, especially when the possible human pose is biased through fixed perspective or some restriction placed on body motion. Our method considers the bias of pose in the target scene and so achieves more accurate domain adaptation (Figure 1). The pose prior of the target scene is generated from various motion capture data sets and environmental knowledge about the scene, and we propose an articulated pose estimation method based on the use of additional pose priors. In experiments, we use the Leeds Sports Poses (LSP) dataset [4] for training and the MPII Cooking Activities dataset [6] for testing, because the videos in the Cooking Activities dataset are captured from a fixed point of view in contrast to those in the LSP dataset which are captured from various perspectives. Using a synthetic pose prior about camera arrangement and without any labeled data about the target domain, our method significantly improves recognition accuracy.

2 Related works

Domain adaptation of articulated pose estimation has been reported in a few papers that examined the estimation accuracy by testing datasets different from that of the training dataset [2]. Although they achieved high accuracy in spite of the different domains because of their part-based approach, they only evaluated versatility of the methods and did not propose any specific method for domain adaptation. On the other hand, Charles et al. approached domain adaptation of pose estimation for people wearing different clothing [7]. Although they retrain the estimator by synthetic training images without any manually labeled data, their method only deals with differences in sleeve length and they ignored the bias of human pose.

Since manually preparing training data is always time-consuming, synthesizing pseudo training data has become a popular technique. With regard to human pose estimation, because the variation in human pose is relatively easy to simulate using motion capture data, learning methods based on such synthetic data are employed in several studies [8, 9]. They synthesize pseudo images or depth maps based on 3D positions of joints, and the resulting trained estimator can be applied to real images. Although our approach also utilizes synthetic data from motion capture, we do not directly synthesize the training images, but rather build pose priors for efficient estimation based on the positions of joints projected onto the camera view because it is difficult to synthesize RGB images covering various appearances from motion data.

3 Basic articulated pose estimator

First, we describe the basic pose estimator based on parts detection by DCNN and its optimization using the skeleton model. To utilize the pose prior of the target domain efficiently, we employ Image Dependent Pairwise Relations (IDPRs) proposed by Chen and Yuille [2] as the detection method. In IDPRs, each joint has several types reflecting the spatial relationships toward neighboring joints and the appearance of each relationship type is trained as a single class. The spatial relationship types are defined discretely using training data, and the multi-class classifiers for each joint and each spatial relationships are trained by DCNNs. The skeleton model for estimation has 18 joints in the upper body and the joints and bones construct a tree structure. The head is taken to be the root node and hands and hips are leaf nodes. We denote the set of joint nodes in the model by $\mathcal{V}$, and the set of connected pairs of joints by $\mathcal{E}$. We use 'neighboring joints' to mean connected joints in the model; 'parent joint' is the root side of connected joints and 'child joint' is the leaf side.

3.1 Parts detection

In IDPRs, the spatial relationships between each neighboring joints are discretized into several types depending on relative positions in the images. Before training the classifier for parts detection, each spatial relationship type is defined based on labeled data. Given the labeled positive images, let $\boldsymbol{d}_{ij}$ be the relative position from joint i to neighbor j. We cluster the relative positions over the training set $\{\boldsymbol{d}_{ij}\}^N$ for each neighboring pair to get K clusters. The K-means algorithm is used for clustering. The type label t_{ij} for each training instance is derived based on its cluster index, and the mean relative position of each pair $\boldsymbol{r}_{ij}^k$ is learnt as the center of cluster k. After clustering, type labels t_{ij} are added to the training dataset.

After deriving type labels, we get a set of labeled patches containing $\boldsymbol{I}(\boldsymbol{l})$: local image patch centered at annotated joint location $\boldsymbol{l}$, $\boldsymbol{c}$: joint labels and $\boldsymbol{m}$: spatial relation types with all its neighbors, from positive images. We use these labeled patches and background patches to train a multi-class DCNN classifier. Output of the classifier is normalised by the softmax function and defined as conditional likelihood function $P(c = i, m_{ij} = t|\boldsymbol{I}(\boldsymbol{l}))$.

3.2 General pose prior

In the combined approach of parts detection and skeleton model, the skeleton model is used to evaluate the plausibility of joint arrangements in terms of as human pose. Although this skeleton model can be considered as prior knowledge of human pose, it is a general pose prior without any bias and so comprehensively covers human poses. Our method also uses the standard skeleton model as a basic pose prior and integrates the domain specific pose prior that suits the domain knowledge. Pictorial Structure (PS) [10] is widely used as a general pose prior in skeleton model application. In PS, the spatial relationships between neighboring joints are defined in a preliminary step and appropriate pose is estimated by optimizing the score function, which penalizes the deviation from standard relative positions. We use a score function similar to PS that is based on classified spatial relationship types about each joint and mean relative position of each type $\boldsymbol{r}_{ij}$.

We optimize the positions of joints by maximizing the following function, which fuses appearance based score by DCNN with arrangement based score by skeleton model:

$$\begin{aligned} F(\boldsymbol{l}, \boldsymbol{t}|\boldsymbol{I}) &= \sum_{i\in\mathcal{V}} U(\boldsymbol{l}_i|\boldsymbol{I}) + \sum_{(i,j)\in\mathcal{E}} \{R_d(\boldsymbol{l}_i, \boldsymbol{l}_j, t_{ij}, t_{ji}) \\ &\quad + R_a(\boldsymbol{l}_i, t_{ij}|\boldsymbol{I}) + R_a(\boldsymbol{l}_j, t_{ji}|\boldsymbol{I})\}, \qquad (1) \end{aligned}$$

where $\boldsymbol{l}$ means the positions of all joints and $\boldsymbol{t}$ means the spatial relation types of all pairs of neighboring joints. Function U, R_d, R_a denotes the appearance likelihood of each joint, spatial likelihood of each pair and appearance likelihood of each spatial relation type of each joint, respectively, as in:

$$\begin{aligned} U(\boldsymbol{l}_i|\boldsymbol{I}) &= w_i^{(U)} \log(P(c = i|\boldsymbol{I}(\boldsymbol{l}_i))), \\ R_d(\boldsymbol{l}_i, \boldsymbol{l}_j, t_{ij}, t_{ji}|\boldsymbol{I}) &= \langle \boldsymbol{w}_{ij}^{(Rd)}, \psi(\boldsymbol{l}_j - \boldsymbol{l}_i - \boldsymbol{r}_{ij})\rangle \\ &\quad + \langle \boldsymbol{w}_{ji}^{(Rd)}, \psi(\boldsymbol{l}_i - \boldsymbol{l}_j - \boldsymbol{r}_{ji})\rangle, \\ R_a(\boldsymbol{l}_i, t_{ij}|\boldsymbol{I}) &= w_{ij}^{(Ra)} \log(P(m_{ij} = t_{ij}|\boldsymbol{I}(\boldsymbol{l}_i))), \end{aligned}$$

where $P(c = i|\boldsymbol{I}(\boldsymbol{l}_i))$ is the output score of DCNN summed up for all spatial types. $\psi()$ is defined as standard quadratic deformation features: $\psi([\delta x, \delta y]) = [\delta x, \delta x^2, \delta y, \delta y^2]^T$ and notation $\langle ., . \rangle$ specifies dot product. $\boldsymbol{r}$ indicates the mean relative position for each spatial relation type, so term R_d evaluates the spatial deformation from the standard relative position. Terms R_a and U evaluate the appearance of each joint as indicated by DCNN. $\boldsymbol{w}^{(U)}, \boldsymbol{w}^{(Rd)}, \boldsymbol{w}^{(Ra)}$ are the weight parameters used to balance the influence of the terms.

Score function of Eq.(1) can be strictly optimized with dynamic programming because the model is a tree structure.

4 Pose prior for domain adaptation

Our idea for applying an estimator trained using a different domain from the target domain is to utilize the bias of poses captured in the target domain. Target domains that have pose bias typically are surveillance videos captured from fixed points of view, and videos of sports that are characterized by limited range of actions. In this paper, the target domain consists of images captured by a static camera like surveillance videos. We assume that there is no labeled data for the target domain, and the only information known is camera position and angle against the floor.

4.1 Synthesizing projected joints

To offset the lack of labeled data, we synthesize pseudo training data using motion capture data to capture the bias in the target domain. Motion capture data is composed of 3D positions of all joints. It is difficult to make RGB images from the data, but the 2D positions in specific camera views can be calculated given the camera parameters. Given the relative position of the camera from the floor on which the target people stand, positions of joints in the image are synthesized by projecting 3D motion capture data to the camera view. Applying this process to comprehensive motion data sets, we obtain the bias of pose in the target camera view.

4.2 Statistical parameters of pose prior

The pose prior of the target domain is represented by the frequency of the spatial relation types defined in Sec.3.1. In particular, the following statistic parameters are calculated based on synthesized pose data.

- The frequency at which the spatial relation type of a joint to its neighboring joint is t.

- The frequency at which the spatial relation type of a joint to its parent joint is t given the spatial relation type of the joint to its children joints.

- The frequency at which the spatial relation type of a joint to its child joint is t given the spatial relation type of the child joint to the joint.

Based on these parameters, the following functions are defined as the likelihood of spatial relation types:

$$\begin{aligned}\phi_a(t|\boldsymbol{t}_i) &= P(m_{ip(i)} = t|\{m_{ik} = t_{ik} : k \in \mathbb{C}(i)\}) \quad (2)\\ \phi_c(t|t_{ji}) &= P(m_{ij} = t|m_{ji} = t_{ji}), \quad (3)\end{aligned}$$

where $p(i)$ means the parent joint of joint i and $\mathbb{C}(i)$ means the set of children joints of joint i. Articulated pose estimation in the target domain is achieved by adding these likelihoods for all joints to basic score function Eq.(1).

4.3 Score function with pose prior

The proposed score function with pose prior is defined as:

$$\begin{aligned}F(\boldsymbol{l}, \boldsymbol{t}|\boldsymbol{I}) &= \sum_{i \in \mathcal{V}} \{U(\boldsymbol{l}_i|\boldsymbol{I}) + J_a(t_{ip(i)}|\{t_{ik} : k \in \mathbb{C}(i)\})\}\\ &\quad + \sum_{(i,j) \in \mathcal{E}} \{R_d(\boldsymbol{l}_i, \boldsymbol{l}_j, t_{ij}, t_{ji}) + R_a(\boldsymbol{l}_i, t_{ij}|\boldsymbol{I})\\ &\quad + R_a(\boldsymbol{l}_j, t_{ji}|\boldsymbol{I}) + J_c(t_{ij}|t_{ji})\}. \quad (4)\end{aligned}$$

Term J_a, which evaluates the likelihood of the combination of spatial relations to neighboring joints and term J_c, which evaluates the connectivity of the spatial relations of neighboring joint pair are added to the basic function Eq.(1). Given the statiscal terms Eq.(2) and Eq.(3), J_a and J_c are defined as:

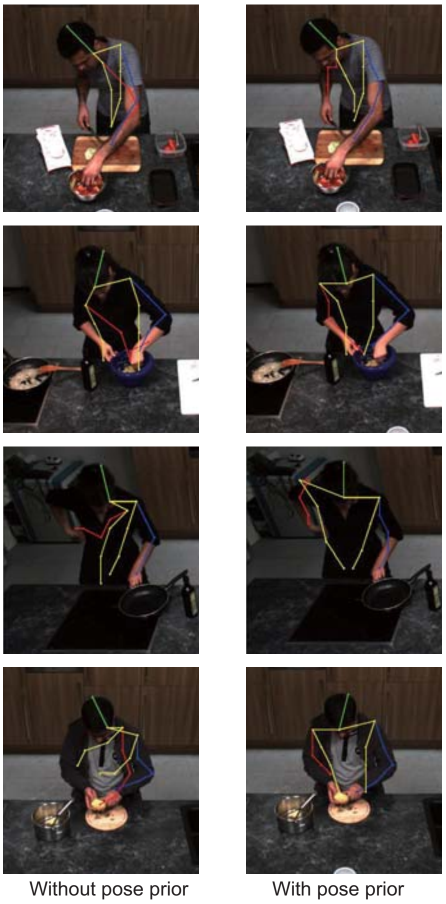

Figure 2. Examples of estimation results. Estimated positions of limbs are indicated by colored lines (green: neck, yellow: torso, red: right arm, blue: left arm).

$$\begin{aligned}J_a(t_{ip(i)}|\{t_{ik} : k \in \mathbb{C}(i)\}) &= w^{(J)} \log(\phi_a(t_{ip(i)}|\boldsymbol{t}_i)),\\ J_c(t_{ij}|t_{ji}) &= w^{(J)} \log(\phi_c(t_{ij}|t_{ji})),\end{aligned}$$

where $w^{(J)}$ is the weight parameter to adjust the influence of the pose prior.

To detect the optimal configuration for each person, we search for the combinations of the location $\boldsymbol{l}$ and type $\boldsymbol{t}$ that maximize the score function of Eq.(4). Since our skeleton model is a tree structure, this can be done efficiently via dynamic programming like [2].

$$(\boldsymbol{l}, \boldsymbol{t}) = \arg\max_{(\boldsymbol{l},\boldsymbol{t})}\{F(\boldsymbol{l}, \boldsymbol{t}|\boldsymbol{I})\}. \quad (5)$$

5 Experiments

We conduct a domain adaptation experiment using 2 datasets from different domains. Specifically, we choose the videos captured by fixed camera as the data for testing assess the bias of pose in the target domain.

5.1 Experimental settings

The Leeds Sports Poses (LSP) dataset [4] is employed as the training dataset. This dataset contains 1000 training images from sport activities with annotated full-body human poses. The poses in this dataset

Table 1. PCP scores on Cooking Activities Dataset about arms. RU: Right upper, LU: Left upper, RF: Right forearm, LF: Left forearm.

$w^{(J)}$	RU	LU	RF	LF
0.01	68.8	74.3	58.4	63.9
0.03	74.2	76.1	60.8	64.5
0.05	76.6	77.7	**62.2**	**65.6**
0.07	77.0	**78.1**	**62.2**	64.6
0.1	**77.4**	76.8	60.8	62.0
w/o prior	64.8	69.9	51.2	58.9

are vary widely and images are captured from various points of view. It is used to train the appearance-based classifier DCNN and define spatial relation types. Moreover, the weight parameters $\boldsymbol{w}^{(U)}, \boldsymbol{w}^{(Ra)}, \boldsymbol{w}^{(Rd)}$ are trained using the validation images by Structured Supported Vector Machine. Although the annotated model in the dataset is for full-body pose, we use only annotations of the upper-body for flexible domain adaptation. The number of types of spatial relations, K, is set to 13. In actual implementation, we use a pre-trained model available at the author's web site [2]. DCNN classifier implementation is based on Caffe framework [11] and the Matlab code made open by Chen and Yuille [2] is used.

The MPII Cooking Activities dataset [6] is employed as the test dataset. This dataset contains videos captured by a static camera with annotated upper-body human poses. Since the camera parameters are not specified in the dataset, we manually defined them according to the videos. The elevation angle is set to $-45°$, and the distance from the foot is set to 5m. To synthesize pseudo pose data, we used the CMU Motion Capture Database [12] as it provides comprehensive motion capture data. For efficient computing, we built pose prior statistics from 20% of all frames in the dataset. With regard to weight parameter $w^{(J)}$, because we did not have the training data to optimize it, we tried several values in the experiment. Since there is only one person at most in every frame of the dataset, after searching for the pose that maximized the score function, only optimal pose computed for each frame is considered as the estimation output. To deal with the scale variation possible given the image patch of joints, we prepared several resized input images for each frame.

5.2 Results of domain adaptation

Table 1 and Figure 2 show the results of the experiment. Estimation accuracy was evaluated by Percentage of Correct Parts (PCP) a standard evaluation metric for articulated pose estimation. To calculate PCP, a estimated limb is considered as correct if both of its joints lie within 50% of the length of the ground-truth limb from annotated joints; PCP is the percentage of frames that are correctly estimated. Because labeling manner of LSP and Cooking Activities are partly different, we use only upper arms and forearms for evaluation. The weight parameter $w^{(J)}$ is configured from 0.01 to 0.1 by reference to the other weight parameters trained with LSP. The table indicates that although the difference of the weight parameter influences the estimation accuracy, the scores are improved by adding pose prior in general. In case of $w^{(J)} = 0.05$, the PCP score is significantly improved from baseline, 11.8% improvement about right upper arm and 11.0% improvement about right forearm. This result supports that synthetic pose prior is helpful for domain adaptation from LSP to Cooking Activities.

Figure 2 shows examples of estimation yielded by the baseline and proposed method ($w^{(J)} = 0.05$). Because of the variety of poses in the LSP dataset, the estimator sometimes outputs unnatural pose and is easily affected by wrong detection of joints. The pose prior suppresses unnatural poses in this domain and offers more accurate pose estimation. For example, the locations of both hands tend to be the same without the prior because they are very similar in appearance.

6 Conclusions

This paper proposes the use of synthetic pose priors for domain adaptation of articulated pose estimation. Especially for videos captured by static cameras, pose priors are easily synthesized and improve the accuracy of recognition in the target domain significantly. Experiments showed that the proposed method attain good results when the estimator, trained with LSP dataset, was applied to the Cooking Activities dataset.

References

[1] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE CVPR*, pp. 1653–1660, 2014.

[2] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Proceedings of NIPS*, pp. 1736–1744, 2014.

[3] L. Sigal, et al. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, Vol. 87, No. 1-2, pp. 4–27, 2010.

[4] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the BMVC*, 2010.

[5] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE CVPR*, pp. 3674–3681, 2013.

[6] M. Rohrbach, et al. A database for fine grained activity detection of cooking activities. In *Proceedings of the IEEE CVPR*, pp. 1194–1201, 2012.

[7] J. Charles, et al. Domain adaptation for upper body pose tracking in signed tv broadcasts. In *Proceedings of the BMVC*, pp. 1–11, 2013.

[8] J. Shotton, et al. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, Vol. 56, No. 1, pp. 116–124, 2013.

[9] G. Shakhnarovich, et al. Fast pose estimation with parameter-sensitive hashing. In *Proceedings of the IEEE ICCV*, pp. 750–757, 2003.

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, Vol. 61, No. 1, pp. 55–79, 2005.

[11] Y. Jia, et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] CMU Motion Capture Database.