Test-time Adaptation for 3D Human Pose Estimation

Sikandar Amin[1,2], Philipp Müller[2], Andreas Bulling[2], and Mykhaylo Andriluka[2,3]

[1]Technische Universität München, Germany
[2]Max Planck Institute for Informatics, Germany
[3]Stanford University, USA

Abstract. In this paper we consider the task of articulated 3D human pose estimation in challenging scenes with dynamic background and multiple people. Initial progress on this task has been achieved building on discriminatively trained part-based models that deliver a set of 2D body pose candidates that are then subsequently refined by reasoning in 3D [1, 4, 5]. The performance of such methods is limited by the performance of the underlying 2D pose estimation approaches. In this paper we explore a way to boost the performance of 2D pose estimation based on the output of the 3D pose reconstruction process, thus closing the loop in the pose estimation pipeline. We build our approach around a component that is able to identify true positive pose estimation hypotheses with high confidence. We then either retrain 2D pose estimation models using such highly confident hypotheses as additional training examples, or we use similarity to these hypotheses as a cue for 2D pose estimation. We consider a number of features that can be used for assessing the confidence of the pose estimation results. The strongest feature in our comparison corresponds to the ensemble agreement on the 3D pose output. We evaluate our approach on two publicly available datasets improving over state of the art in each case.

1 Introduction and related work

In this paper we consider the task of articulated 3D human pose estimation from multiple views. We focus on the setting with uncontrolled environment, dynamic background and multiple people present in the scene, which is more complex and general compared to the motion capture studio environments often considered in the literature [7, 13, 17]. One of the key challenges in that setting is that the appearance of people is more diverse, and simple means of representing observations based on background subtraction are not applicable due to the presence of multiple people and interactions between people and scene objects. Inspired by recent results in 2D pose estimation [3, 18], several approaches have proposed to build upon and adapt these results for pose estimation in 3D [1, 4, 5, 12]. In these approaches 2D detectors are either used to model the likelihood of the 3D pose [5, 12], or provide a set of proposals for positions of body joints that are subsequently refined by reasoning in 3D [1, 4]. Improving the 2D pose estimation performance is thus crucial for each of these methods. Towards this goal we propose an approach to tune the 2D pose estimation component at test time.

Generally, one would expect that the pose estimation results should improve if one is continuously observing the same scene with the same human subjects, as one would be able to learn more specific appearance models than is possible in the general case.

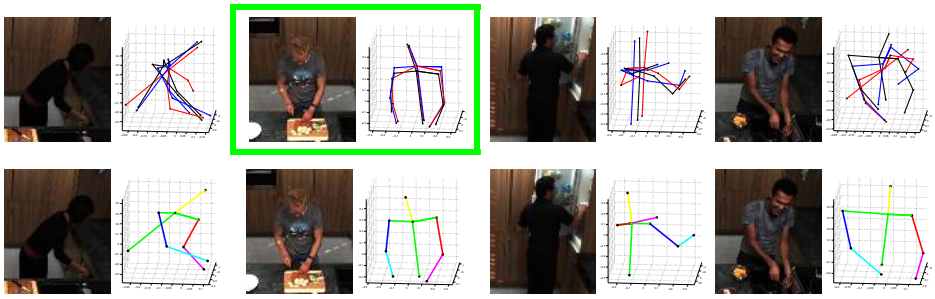


Fig. 1: Overview of our approach. *Top row:* In the first stage we estimate 3D poses of people in each frame with an ensemble of multi-view pictorial structures models. Output of the models from the ensemble is shown in blue, red and black. We select highly confident key-frames either based on (1) agreement between models in the ensemble, or (2) using a classifier trained on features computed from these outputs. The green bounding box indicates a selected key-frame. *Bottom row*: Output of our final model that incorporates evidence from the keyframes.

However, with a few exceptions [6, 14], this idea is rarely explored in the literature, likely because it is unclear how to robustly estimate the person-specific appearance in the presence of noise in the pose estimation. Various approaches for estimating the confidence of the pose prediction have been considered in the literature, ranging from models that are discriminatively trained for both detection and pose estimation [18] to specialized methods that estimate confidence based on the combination of features as a post-processing step [11]. In this paper we follow the direction similar to [11] but also employ features based on the 3D pose reconstruction, which we find to be highly effective for filtering out incorrect pose estimates. Overall we make the following contributions. As a main contribution of this paper we propose a new approach for articulated 3D human pose estimation that builds on the multi-view pictorial structures model [1], and extends it to adapt to observations available at test time. Our approach has an interesting property that it operates on the entire test set, making use of the evidence available in all test images. This is in contrast to prior works [1, 4, 5, 12] that typically operate on single-frames only or are limited to temporal smoothness constraints which are effective only in a small temporal neighborhood of each frame [2, 16]. As a second contribution we evaluate two approaches to assess the accuracy of 3D pose estimation. The first is to train a discriminative model based on various pose quality features as in [11], and the second is to consider the agreement of an ensemble of several independently trained models on the 3D pose output. An interesting finding of our evaluation is that pose agreement alone performs on-par or better than the discriminatively trained confidence predictor. The combination of both approaches further improves the results.

Overview of our approach. In this paper we build on the multi-view pictorial structures approach proposed in [1]. This approach first jointly estimates projections of each body joint in each view, and then recovers 3D pose by triangulation. We explore two mechanisms for improving the performance of the multi-view pictorial structures model. Both of them are based on the observation that 3D pose reconstruction provides strong cues for identification of highly confident pose estimation hypotheses (= key-frames) at test time (see Fig. 1 for a few examples). We explore two ways to take advantage of such

key-frame hypotheses. We either directly use them as additional training examples in order to adapt the 2D pose estimation model to the scene at hand, or we extend the pictorial structures model with an additional term that measures appearance similarity to the key-frames. As we show in the experiments both mechanisms considerably improve the pose estimation results. In the following we first introduce the multi-view pictorial structures model and then describe our extensions.

2 Multi-view pictorial structures

The pictorial structures model represents a body configuration as a collection of rigid parts and a set of pairwise part relationships [8, 10]. We denote a part configuration as $L = \{l_i | i = 1, \ldots, N\}$, where $l_i = (x_i, y_i, \theta_i)$ corresponds to the image position and absolute orientation of each part. Assuming that the pairwise part relationships have a tree structure the conditional probability of the part configuration L given the image evidence I factorizes into a product of unary and pairwise terms:

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(l_n; I) \cdot \prod_{(i,j) \in E} f_{ij}(l_i, l_j). \tag{1}$$

where $f_n(l_n; I)$ is the likelihood term for part n, $f_{ij}(l_i, l_j)$ is the pairwise term for parts i and j and Z is a partition function.

Multi-view model: Recently [1, 5] have extended this approach to the case of 3D human pose estimation from multiple views. In the following we include the concise summary of the multiview pictorial structures model that we use in our experiments and refer the reader to the original paper [1] for more details.

The multiview pictorial structures approach proposed by [1] generalizes the single-view case by jointly reasoning about the projections of body parts in each view. Let L_v denote the 2D body configuration and I_v the image observations in view v. Multiview constraints are modeled as additional pairwise factors in the pictorial structures framework that relate locations of the same joint in each view. The resulting multiview pictorial structures model corresponds to the following decomposition of the posterior distribution:

$$p(L_1, ..., L_V | I_1, ..., I_V) = \frac{1}{Z} \prod_{v} f(L_v; I_v) \prod_{(a,b)} \prod_{n} f_n^{app}(l_n^a, l_n^b; I_a, I_b) f_n^{cor}(\mathbf{l}_n^a, \mathbf{l}_n^b), \tag{2}$$

where $\{(a, b)\}$ is the set of all view-pairs, $\mathbf{l}_n^v$ represents the image position of part n in view v, in contrast l_n^v in addition to image position also includes the absolute orientation, $f(L_v; I_v)$ are the single-view factors for view v which decompose into products of unary and pairwise terms according to Eq. 1, and f_n^{app} and f_n^{cor} are multiview appearance and correspondence factors for part n. The inference is done jointly across all views. The 2D pose estimation resutls are then triangulated to reconstruct the final 3D pose.

3 Test-time adaptation

Our approach to test-time adaptation is composed of two stages. In the first stage we mine confident pose estimation examples from the test data. These examples are then used in the second stage to improve the pose estimation model. We now describe these stages in detail.

3.1 Confident examples mining

The objective of the first stage is to identify the test examples for which the initial model succeeded in correctly estimating the body pose. To that end, we consider two methods to assess the accuracy of pose estimation.

3D pose agreement. In the first method we proceed by training an ensemble of M multiview pictorial structure models. Each model in the ensemble is trained on a disjoint subset of the training set. In addition we also train a reference model using all training examples. The rationale behind this procedure is that the reference model will typically perform better than ensemble models as it is trained on more examples. Ensemble models will in turn provide sufficient number of independent hypothesis in order to assess the prediction accuracy. At test time we evaluate the agreement between the pose hypotheses estimated by ensemble models and the reference model. Given the estimated 3D poses from all models, we define the pose agreement score s_{pa} as:

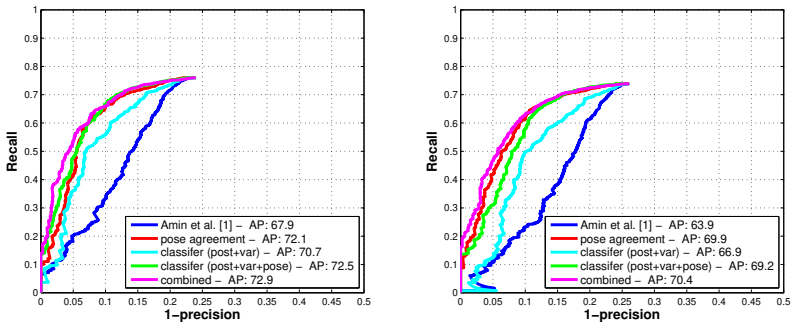
$$s_{pa} = \exp\left(-\frac{\sum_n \sum_m \|\mathbf{x}_n^m - \hat{\mathbf{x}}_n\|_2^2}{N}\right), \tag{3}$$

where $\mathbf{x}_n^m$ represents the 3D position of part n estimated with the ensemble model $m \in \{1, ..., M\}$, and $\hat{\mathbf{x}}_n$ is the location of part n estimated with the reference model.

Pose classification. As a second method we train a discriminative AdaBoost classifier to identify correct 3D pose estimates. The classifier is trained using the following features:

1. *3D pose features.* These features encode the plausible 3D poses and correspond to the torso, head and limb lengths, distance between shoulders, angles between upper and lower limbs, and angles between head and shoulder parts.
2. *Prediction uncertainty features.* We encode the uncertainty in pose estimation by computing the L2 norm of the covariance matrix corresponding to the strongest mode in the marginal posterior distribution of each part in each view. This is the same criteria as used for component selection in [1] and is similar to the features used in [11].
3. *Posterior.* As a separate feature we also include the value of the posterior distribution corresponding to the estimated pose that is given by Eq. 2.

We concatenate these three types of features to produce a combined feature vector of the size $20 + 2VN$ for the upper-body and $26 + 2VN$ for the full-body case, where V is the number of views and N is the number of parts in the pictorial structures model. We rely on a disjoint validation set for training of the pose classifier. We consider 3D poses



(a) Upper Arms (b) Lower Arms

Fig. 2: MPIICooking dataset: Key-frame selection using score based on the posterior marginal of the model from [1] (blue) compared to variants of our approach. Best result corresponds to combination of ensemble agreement and pose classification scores given by Eq. 5 (magenta).

with all body parts estimated correctly to be positive examples and all other examples as negative. Body part is considered to be correctly estimated if both of its endpoints are within 50% of the part length from their ground-truth positions. The classifier score corresponding to the m^{th} pictorial structure model $s_{abc,m}$ is given by the weighted sum of the weak single-feature classifiers $h_{m,t}$ with weights $\alpha_{m,t}$ learned using AdaBoost:

$$s_{abc,m} = \frac{\sum_t \alpha_{m,t} h_{m,t}(x_m)}{\sum_t \alpha_{m,t}} \tag{4}$$

Combined approach. Finally, we consider a weighted combination of agreement and classification scores:

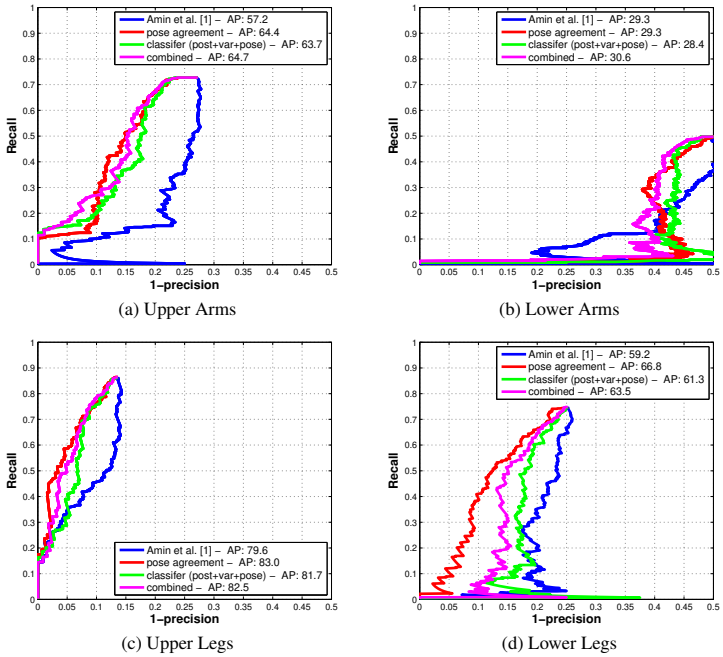
$$s_{comb} = s_{pa} + \sum_m w_m s_{abc,m}, \tag{5}$$

where the weights of the classifier scores are given by $w_m = \frac{\sum_{\hat{m} \neq m} \phi_{\hat{m}}}{(M-1)\sum_{\hat{m}} \phi_{\hat{m}}}$, and $\phi_{\hat{m}}$ are given by the training time mis-classification error. As we demonstrate in section 4 such combination improves results over using each approach individually.

3.2 2D model refinement

At test-time we choose 10% of the highest scoring pose hypotheses according to one of the scoring methods described in Sec. 3.1 and denote them as key-frames. We investigate two different avenues to use key-frames in order to improve pose estimation performance.

Retraining: We retrain the discriminative part classifiers by augmenting the training data with part examples from the key-frames and $n = 5$ of their nearest neighbors, mined from the entire test set. To compute nearest neighbors we encode each part hypothesis using shape context features sampled on the regular grid within the part bounding box and bounding box color histogram. The nearest neighbors are then found using euclidean distance in this feature space. For the rest of the paper, we refer to this approach as **RT**.



(a) Upper Arms (b) Lower Arms

(c) Upper Legs (d) Lower Legs

Fig. 3: Shelf dataset: Key-frame selection using score based on the posterior marginal of the model from [1] (blue) compared to variants of our approach.

Appearance similarity: We introduce additional unary term for each part in the pictorial structures model that encourages similarity to key-frames. The similarity term for part n is given by:

$$f_{SIM}(l_n; I) = \exp\left(-\frac{\min_j \|e(l_n) - e(\mathbf{a}_{nj})\|_2^2}{2 * \sigma_n^2}\right), \tag{6}$$

where l_n is image position and absolute orientation of the part hypothesis, $\mathbf{a}_{nj}$ is a hypothesis for part n from the j-th key-frame, and $e(l_n)$ corresponds to shape-context and color features extracted at l_n. The variance σ_n^2 is estimated based on the euclidean distances between all feature vectors corresponding to part n in the training set. We refer to this approach as **SIM** later in text.

4 Experiments

Datasets: Our aim is to analyze the performance of our proposed approach in challenging settings with dynamic background and a variety of subjects. Therefore, we evaluate our approach on MPII Cooking [15] and the Shelf [4] datasets. Both of these datasets have been recently introduced for the evaluation of articulated human pose estimation from multiple views.

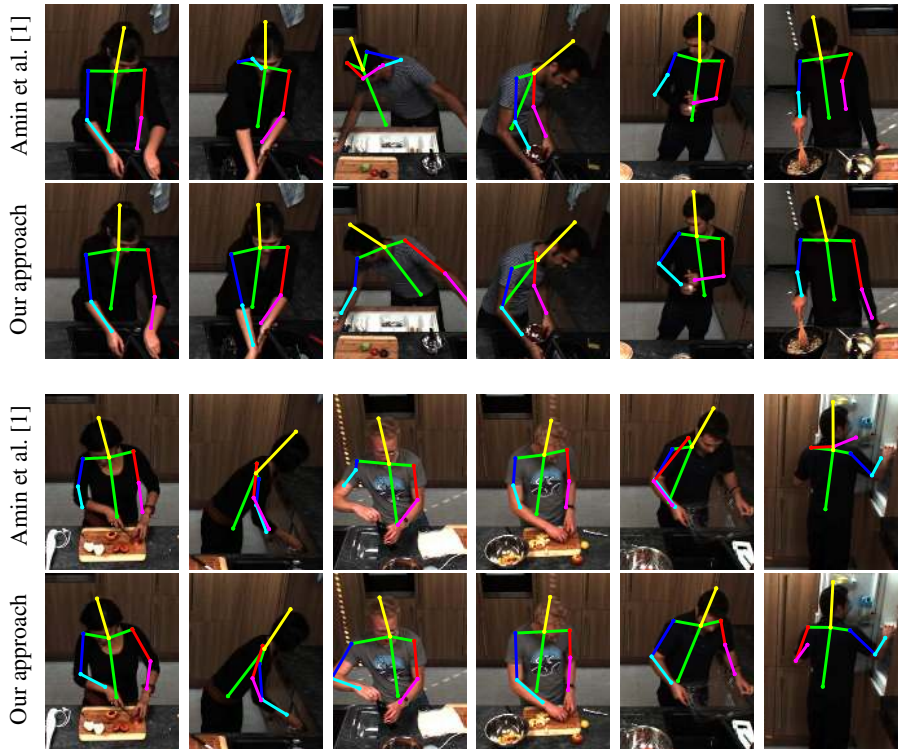


Fig. 4: Examples of pose estimation results obtained with our approach and comparison to the state-of-the-art approach of Amin et al. [1] on the MPII Cooking dataset.

MPII Cooking: The dataset was originally recorded for the task of fine grained activity recognition of cooking activities, and has been later used in [1] to benchmark the performance of 3D pose estimation. This evaluation dataset consists of 11 subjects with non-continuous images and two camera views. The training set includes 4 subjects and 896 images and the test set includes 7 subjects and 1154 images.

Shelf dataset: This dataset has been introduced in [4] and is focused on the task of multiple human 3D pose estimation. The dataset depicts up to 4 humans interacting with each other while performing an assembly task. The Shelf dataset provides 668 and 367 annotated frames for training and testing respectively. For every frame each fully visible person is annotated in 3 camera views.

In our evaluation we rely on the standard train/test split and evaluation protocols as used by the original publications. As described in section 3, we split the training set in multiple parts to train an ensemble of pose estimation models.

Key-frames Analysis: We analyze the performance of our key-frame detection procedure using recall-precision curves. The results are shown in Fig. 2 and 3. In this analysis we omit the torso and head body parts as they are almost perfectly localized by all approaches. For all other parts, which are smaller in size hence more susceptible to noise, we observe that directly using the marginal posterior of the PS model as pose confi-

Fig. 5: Examples of pose estimation failures on the MPIICooking dataset.

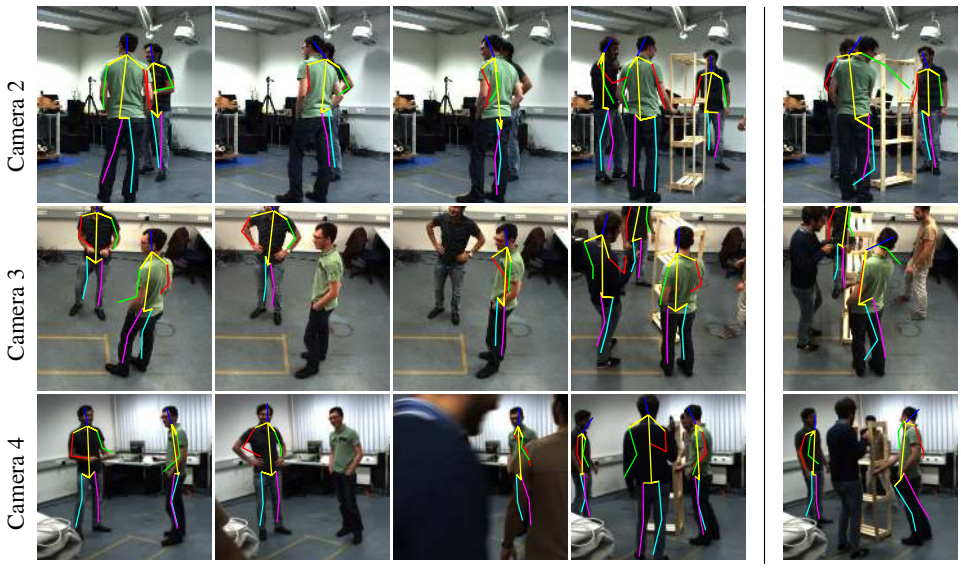


Fig. 6: Examples of pose estimation results of our approach on the Shelf dataset. Last column shows an example of the failure case.

dence leads to poor results (blue curve in Fig. 2 and 3). On the MPII Cooking dataset we get AP of 67.9/63.9 for upper/lower arms respectively using this marginal posterior as confidence measure. On the other hand, training a classifier with posterior and variance features improves the AP to 70.7/66.9. The performance of the classifier increases further to 72.5/69.2 when we extend the feature vector with 3D pose cues. This result underlines the importance of 3D pose features for classification. Interestingly, the results of the pose agreement approach of Eq. 3 i.e. 72.1/69.9, suggest that this measure alone is almost equally effective. Combining both (classifier & pose agreement) scores as in Eq. 5, we step up the average precision of the detected key-frames to 72.9/70.4. These results show the importance of different levels of features in the process of extracting key-frames with high confidence.

The recall-precision curve for the Shelf dataset are shown in Fig. 3. Although, here the detection of lower arms is significantly more difficult, but the AP results for both upper and lower arms are in line with the MPII Cooking AP values. The AP values are slightly worse for the classifier score compared to pose agreement. For the legs, pose agreement alone outperforms the combined score of Eq. 5 i.e., 83.0/66.8 as compared to 82.5/63.5 for upper/lower legs. The reason for this behavior is the significantly worse

Model	Torso	Head	upper arm r	upper arm l	lower arm r	lower arm l	All
Cam-1							
Amin et al. [1]	92.9	89.4	72.6	79.4	68.8	76.8	80.0
our (RT)	95.2	93.8	**75.3**	**83.4**	73.6	80.9	83.6
our (SIM)	**95.8**	92.5	74.0	82.6	73.5	82.2	83.4
our (RT+SIM)	**95.8**	**94.0**	74.8	83.3	**74.2**	**82.3**	**84.0**
Cam-2							
Amin et al. [1]	91.1	92.4	75.4	76.7	72.9	74.7	80.5
our (RT)	92.1	95.5	79.2	82.8	76.9	78.8	84.2
our (SIM)	92.4	**96.2**	**79.5**	81.6	**77.1**	79.6	84.4
our (RT+SIM)	**92.6**	**96.2**	78.9	**83.3**	**77.1**	**79.7**	**84.7**

Table 1: MPII Cooking: accuracy measured using percentage of correct parts (PCP) score [9]. We compare the model of Amin et al. [1] with variants of our approach. *RT* stands for model retraining, *SIM* stands for a model augmented with similarity factors.

	Belagiannis et al. [4]	our (RT)	our (SIM)	our (RT+SIM)
Actor1	66	68.5	**72.1**	72.0
Actor2	65	67.2	69.4	**71.3**
Actor3	83	83.9	84.9	**85.7**
Average	71.3	74.4	77.0	**77.3**

Table 2: Shelf dataset: accuracy measured using 3D PCP score. We compare the model of Belagiannis et al. [4] with variants of our approach. *RT* stands for model retraining, *SIM* stands for a model augmented with similarity factors.

performance of the classifier of Eq. 4. This result suggests the need to learn weights when combining different scores as in Eq. 5. This we will investigate in future work. For the Shelf dataset the training and test splits contain the same subjects. Moreover, the training data splits for ensemble of multiview pictorial structure models also contain the same subjects. This explains the higher preformance of the pose agreement cue compared to the classifier output.

Pose estimation results: Here we discuss the improvement we achieve in pose estimation performance when we incorporate these key-frames for 2D model refinement as discussed in Section 3.2. We use score of the combined approach s_{comb} to select the key-frames. Following [1, 15, 16], we use body-joints instead of limbs as parts in the pictorial structures model. This approach is commonly referred to as flexible pictorial structures model (FPS).

MPII Cooking: We use the standard pose configuration, i.e., 10 upperbody parts as introduced in [15]. Amin et al. [1] reports the percentage of correct parts (PCP) for the 2D projections per camera for this dataset. First, we evaluate our proposed retraining approach (RT) to improve overall pose estimation accuracy by adapting the model to test scene specific settings. We show the PCP results for MPII Cooking in Table 1. Our RT approach achieves 83.6/84.2 overall PCP and shows improvement for all individual parts. This improvement can be attributed to the fact that retraining the model

including the mined examples can learn the person/scene specific features. The other approach (SIM) which involves adding a new unary term, based on the feature similarity, to the pictorial structures model also achieves competitve results, i.e., 83.4/84.4 overall PCP. The improvement in this case is more pronounced on the lower arms compared to retraining the model. Furthermore, we also evaluate the combination of the two approaches RT+SIM. In this approach, along with retraining the part classifiers using the appearance features from the key-frames we also introduce the appearance similarity based unary term f_{SIM} to the multiview pictorial structures framework. The results in Table 1 show that this works best because it combines the benefits of both approaches and results in stronger unaries for the part hypotheses. We illustrate some example improvements of our approach in Fig. 4, as compared to [1] on MPII Cooking dataset. Fig. 5 demonstrates some typical failure cases of our approach.

Shelf dataset: We use a full body model with 14 parts in this case as described in the original paper [4]. The accuracy of the approach from [4] is bounded by the performance of the 2D part detectors. Their model is unable to recover once the 2D part detector fails to fire in the first stage. On the other hand, as our approach is able to utilize the test scene specific information, we achieve far better results in terms of PCP values. Table 2 shows the 3D PCP values in comparison to the recent results of [4]. Our first approach, i.e., retraining the model RT, outperforms the approach from [4] by 3% PCP. Interestingly, we get 77.0 PCP with our second approach SIM which involves model inference using an extra similarity term and it outperforms our RT approach by further 2.6%. This result can be explained by the fact that all dimensions of the appearance feature vector are considered equally in this approach. There exist some features which do not perform well during the feature selection process in AdaBoost when learnt together with the training set. Still, they contain similarity information for the test examples when compared against the mined key-frames in terms of euclidean distance for the complete feature vector. Further gain of 0.3 PCP is obtained by combining RT with SIM. Some examples of qualitative results on Shelf dataset are depicted in Fig. 6.

5 Conclusion

In this paper we proposed an approach to 3D pose estimation that adapts to the input data available at test time. Our approach operates by identifying frames in which poses can be predicted with high confidence and then uses them as additional training examples and as evidence for pose estimation in other frames. We analyzed two strategies for finding confident pose estimates: discriminative classification and ensemble agreement. Best results are achieved by combining both strategies. However, ensemble agreement alone already improves considerably over the confidence measure based on the pictorial structures output. We have shown the effectiveness of our approach on two publicly available datasets. In the future we plan to generalize our approach to multiple rounds of confident examples mining, and will explore other approaches for automatic acquisition of training examples from unlabeled images.

Acknowledgements. This work has been supported by the Max Planck Center for Visual Computing and Communication.

References

1. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3d human pose estimation. In: BMVC (2013)
2. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR (2010)
3. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. IJCV (2011)
4. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: CVPR (2014)
5. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: CVPR (2013)
6. Eichner, M., Ferrari, V.: Appearance sharing for collective human pose estimation. In: ACCV (2012)
7. El Hayek, A., Stoll, C., Hasler, N., Kim, K.i., Seidel, H.P., Theobalt, C.: Spatio-temporal motion tracking with unsynchronized cameras. In: CVPR (2012)
8. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV (2005)
9. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)
10. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computers C-22(1), 67–92 (1973)
11. Jammalamadaka, N., Zisserman, A., Eichner, M., Ferrari, V., Jawahar, C.V.: Has my algorithm succeeded? an evaluator for human pose estimators. In: ECCV (2012)
12. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: BMVC (2013)
13. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley MHAD: A comprehensive multimodal human action database. In: WACV (2013)
14. Ramanan, D., Forsyth, D.A., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. In: CVPR (2005)
15. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: CVPR (2012)
16. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR (2011)
17. Sigal, L., Balan, A., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV 87(1-2) (2010)
18. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)