Iterative Greedy Matching for 3D Human Pose Tracking from Multiple Views

Julian Tanke and Juergen Gall
{tanke, gall}@iai.uni-bonn.de

University of Bonn

Abstract. In this work we propose an approach for estimating 3D human poses of multiple people from a set of calibrated cameras. Estimating 3D human poses from multiple views has several compelling properties: human poses are estimated within a global coordinate space and multiple cameras provide an extended field of view which helps in resolving ambiguities, occlusions and motion blur. Our approach builds upon a real-time 2D multi-person pose estimation system and greedily solves the association problem between multiple views. We utilize bipartite matching to track multiple people over multiple frames. This proofs to be especially efficient as problems associated with greedy matching such as occlusion can be easily resolved in 3D. Our approach achieves state-of-the-art results on popular benchmarks and may serve as a baseline for future work.

1 Introduction

3D human pose tracking has applications in surveillance [40] and analysis of sport events [7, 23]. Most existing approaches [19, 21, 25–27, 33, 38, 28, 29] address 3D human pose estimation from single images while multi-view 3D human pose estimation [7, 23, 3, 4, 12] remains less explored, as obtaining and maintaining a configuration of calibrated cameras is difficult and costly. However, in sports or surveillance, calibrated multi-camera setups are available and can be leveraged for accurate human pose estimation and tracking. Utilizing multiple views has several obvious advantages over monocular 3D human pose estimation: ambiguities introduced by foreshortening as well as body joint occlusions or motion blurs can be resolved using other views. Furthermore, human poses are estimated within a global coordinate system when using calibrated cameras.

In this work we propose an iterative greedy matching algorithm based on epipolar geometry to approximately solve the k-partite matching problem of multiple human detections in multiple cameras. To this end we utilize a real-time 2D pose estimation framework and achieve very strong results on challenging multi-camera datasets. The common 3D space proves to be very robust for greedy tracking, resulting in a very efficient and well-performing algorithm. In contrast to previous works [7, 23, 34, 13], our approach does not discretize the solution space but combines triangulation with an efficient pose association approach across camera views and time. Furthermore, our approach does not utilize individual shape models for each person [26].
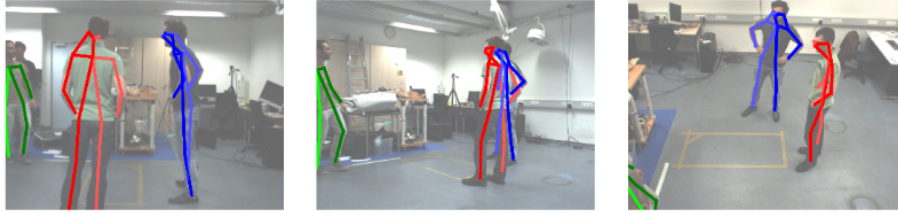
Fig. 1. Qualitative results on the Shelf [3] dataset.

We make the following contributions: (i) we present a greedy approach for 3D multi-person tracking from multiple calibrated cameras and show that our approach achieves state-of-the-art results. (ii) We provide extensive experiments on both 3D human pose estimation and on 3D human pose tracking on various multi-person multi-camera datasets.

2   Related Work

Significant progress has been made in pose estimation and pose tracking in recent years [8, 11, 20, 39] and our model is built on advancements in the field of 2D multi-person pose estimation [8, 9, 15, 17, 24, 31, 36, 39]. For instance, part affinity fields [8] are 2D vector fields that represent associations between body joints which form limbs. It utilizes a greedy bottom-up approach to detect 2D human poses and is robust to early commitment. Furthermore, it decouples the runtime complexity from the number of people in the image, yielding real-time performance.

There is extensive research in monocular 3D human pose estimation [19, 21, 25, 27, 33, 38, 28, 29]. For instance, Martinez et al. [27] split the problem of inferring 3D human poses from single images into estimating a 2D human pose and then regressing the 3D pose on the low-dimensional 2D representation. Though 3D human pose estimation approaches from single images yield impressive results they do not generalize well to unconstrained data.

While multiple views are used in [34, 35] to guide the training for monocular 3D pose estimation, there are also approaches that use multiple views for inference. A common technique to estimate a single 3D human pose from multiple views is to extend the well-known pictorial structure model [14] to 3D [2, 5, 7, 23, 34]. Burenius et al. [7] utilize a 2D part detector based on the HOG-descriptor [10] while Kazemi et al. [23] use random forests. Pavlakos et al. [34] outperform all previous models by utilizing the stacked hourglass network [32] to extract human joint confidence maps from the camera views. However, these models have to discretize their solution space resulting in either a very coarse result or a very large state space making them impractical for estimating 3D poses of multiple people. Furthermore, they restrict their solution space to a 3D bounding volume around the subject which has to be known in advance. Estimating multiple humans from multiple views was first explored by Belagiannis
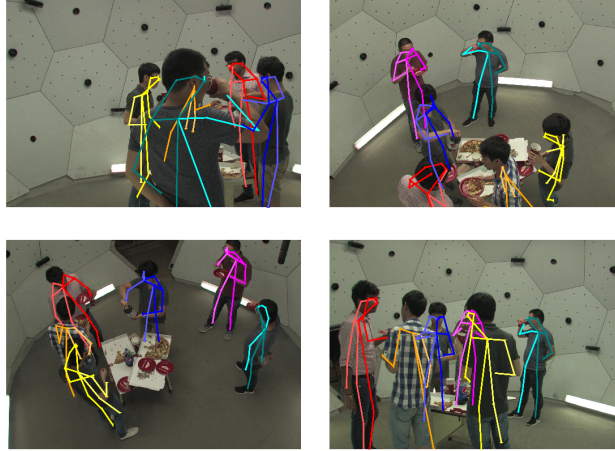
Fig. 2. Challenging 3D reconstruction of 6 persons in the *CMU Panoptic Dataset* [22] with significant occlusion and partial visibility of persons.

et al. [3, 4]. Instead of sampling from all possible translations and rotations they utilize a set of 3D body joint hypotheses which were obtained by triangulating 2D body part detections from different views. However, these methods rely on localizing bounding boxes using a person tracker for each individual in each frame to estimate the number of persons that has to be inferred from the common state space. This will work well in cases where individuals are completely visible in most frames but will run into issues when the pose is not completely visible in some cameras as shown in Figure 2. A CNN-based approach was proposed by Elhayek et al. [12] where they fit articulated skeletons using 3D sums of Gaussians [37] and where body part detections are estimated using CNNs. However, the Gaussians and skeletons need to be initialized beforehand for each actor in the scene, similar to [26]. Fully connected pairwise conditional random fields [13] utilize approximate inference to extract multiple human poses where DeeperCut [18] is used as 2D human pose estimation model. However, the search space has to be discretized and a fully connected graph has to be solved, which throttles inference speed. Our approach does not suffer from any of the aforementioned drawbacks as our model works off-the-shelf without the need of actor-specific body models or discretized state space and uses an efficient greedy approach for estimating 3D human poses.

3   Model

Our model consists of two parts: First, 3D human poses are estimated for each frame. Second, the estimated 3D human poses are greedily matched into tracks which is described in Section 3.2. To remove outliers and to fill-in missing joints
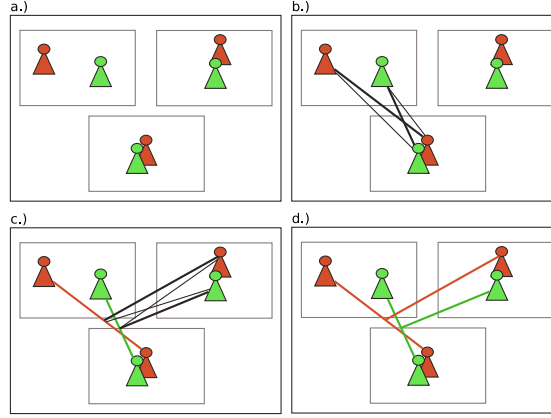
Fig. 3. Estimating multiple people from multiple views can be formulated as k-partite graph partitioning where 2D human pose detections must be associated across multiple views. We employ a greedy approach to make the partitioning tractable. Given a set of 2D human pose detections on multiple views (a) we greedily match all detections on two images (b) where the weight between two detections is defined by the average epipolar distance of the two poses. Other views are then integrated iteratively where the weight is the average of the epipolar distance of the 2D detections in the new view and the already integrated 2D detections (c). 2D detections with the same color represent the same person.

in some frames, a simple yet effective smoothing scheme is applied, which is also discussed in Section 3.2.

3.1   3D Human Pose Estimation

First, 2D human poses are extracted for each camera separately. Several strong 2D multi-person pose estimation [8, 9, 15, 17, 24, 31, 36, 39] models have been proposed but in our baseline we utilize OpenPose [8] as it is well established and offers real-time capabilities. We denote the 2D human pose estimations as

$$\left\{h_{i,k}\right\}_{i\in[1,N]}^{k\in[1,K_i]} \tag{1}$$

where N is the number of calibrated cameras and K_i the number of detected human poses for camera i.

In order to estimate the 3D human poses from multiple cameras, we first associate the detections across all views as illustrated in Figure 3. We denote the associated 2D human poses as $\mathcal{H}$ where $|\mathcal{H}|$ is the number of detected persons and $\mathcal{H}_m = \{h_{i,k}\}$ is the set of 2D human poses that are associated to person m. Once the poses are associated, we estimate the 3D human poses for all detected persons m with $|\mathcal{H}| > 1$ by triangulating the 2D joint positions.

For the association, we select camera $i = 1$ as starting point and choose all 2D human pose detections $h_{1,k}$ in this camera view as person candidates,
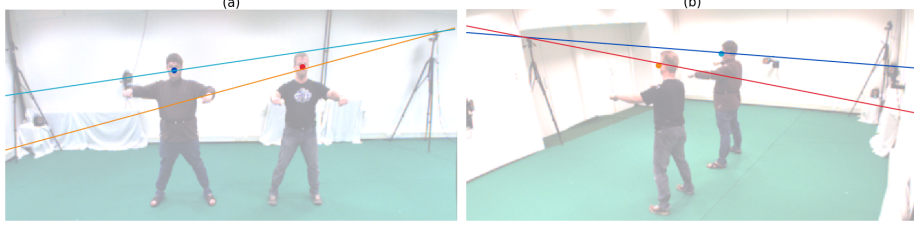
Fig. 4. Epipolar lines for two camera views of the UMPM Benchmark [1]. The blue and the red dot in image (a) are projected as blue (red) epipolar lines in the second image (b) while the orange and light-blue dot from image (b) are projected onto image (a).

i.e., $\mathcal{H} = \{\{h_{1,k}\}\}$. We then iterate over the other cameras and greedily match their 2D detections with the current list of person candidates $\mathcal{H}$ using bi-partite matching [30].

The cost for assigning a pose $h_{i,k}$ to an existing person candidate $\mathcal{H}_m$ is given by

$$\Phi(h_{i,k}, \mathcal{H}_m) = \frac{1}{|\mathcal{H}_m||J_{kl}|} \sum_{h_{j,l} \in \mathcal{H}_m} \sum_{\iota \in J_{kl}} \phi(h_{i,k}(\iota), h_{j,l}(\iota)) \tag{2}$$

where $h_{i,k}(\iota)$ denotes the 2D pixel location of joint ι of the 2D human pose $h_{i,k}$ and J_{kl} is the set of joints that are visible for both poses $h_{i,k}$ and $h_{j,l}$. Note that the 2D human pose detections might not contain all J joints due to occlusions or truncations. The distance between two joints in the respective cameras is defined by the distance between the epipolar lines and the joint locations:

$$\phi(p_i, p_j) = |p_j^T F^{i,j} p_i| + |p_i^T F^{j,i} p_j| \tag{3}$$

where $F^{i,j}$ is the fundamental matrix from camera i to camera j. Figure 4 shows the epipolar lines for two joints.

Using the cost function $\Phi(h_{i,k}, \mathcal{H}_m)$, we solve the bi-partite matching problem for each image i:

$$X^* = \operatorname*{argmin}_{X} \sum_{m=1}^{|\mathcal{H}|} \sum_{k=1}^{K_i} \Phi(h_{i,k}, \mathcal{H}_m) X_{k,m} \tag{4}$$

where

$$\sum_k X_{k,m} = 1 \ \forall m \quad \text{and} \quad \sum_m X_{k,m} = 1 \ \forall k.$$

$X_{k,m}^* = 1$ if $h_{i,k}$ is associated to an existing person candidate $\mathcal{H}_m$ and it is zero otherwise. If $X_{k,m}^* = 1$ and $\Phi(h_{i,k}, \mathcal{H}_m) < \theta$, the 2D detection $h_{i,k}$ is added to $\mathcal{H}_m$. If $\Phi(h_{i,k}, \mathcal{H}_m) \geq \theta$, $\{h_{i,k}\}$ is added as hypothesis for a new person to $\mathcal{H}$. Algorithm 1 summarizes the greedy approach for associating the human poses across views.

Result: Associated 2D poses $\mathcal{H}$

$\mathcal{H} := \big\{\{h_{1,k}\}\big\}$;

for *camera* $i \leftarrow 2$ **to** N **do**

    for *pose* $k \leftarrow 1$ **to** K_i **do**

        for *hypothesis* $m \leftarrow 1$ **to** $|\mathcal{H}|$ **do**

            $C_{k,m} = \Phi(h_{i,k}, \mathcal{H}_m)$;

        end

    end

    $X^* = \underset{X}{\operatorname{argmin}} \sum_{m=1}^{|\mathcal{H}|} \sum_{k=1}^{K_i} C_{k,m} X_{k,m}$;

    for k, m ***where*** $X^*_{k,m} = 1$ **do**

        if $C_{k,m} < \theta$ **then**

            $\mathcal{H}_m = \mathcal{H}_m \bigcup \{h_{i,k}\}$;

        else

            $\mathcal{H} = \mathcal{H} \bigcup \big\{\{h_{i,k}\}\big\}$;

        end

    end

end

$\mathcal{H} = \mathcal{H} \setminus \mathcal{H}_m \ \forall m$ **where** $|\mathcal{H}_m| = 1$;

Algorithm 1: Solving the assignment problem for multiple 2D human pose detections in multiple cameras. $\Phi(h_{i,k}, \mathcal{H}_m)$ (2) is the assignment cost for assigning the 2D human pose $h_{i,k}$ to the person candidate $\mathcal{H}_m$. X^* is a binary matrix obtained by solving the bi-partite matching problem. The last line in the algorithm ensures that all hypotheses that cannot be triangulated are removed.

3.2   Tracking

For tracking, we use bipartite matching [30] similar to Section 3.1. Assuming that we have already tracked the 3D human poses until frame $t-1$, we first estimate the 3D human poses for frame t as described in Section 3.1. The 3D human poses of frame t are then associated to the 3D human poses of frame $t-1$ by bipartite matching. The assignment cost for two 3D human poses is in this case given by the average Euclidean distance between all joints that are present in both poses. In some cases, two poses do not have any overlapping valid joints due to noisy detections or truncations. The assignment cost is then calculated by projecting the mean of all valid joints of each pose onto the xy-plane, assuming that the z-axis is the normal of the ground plane, and taking the Euclidean distance between the projected points. As long as the distance between two matched poses is below a threshold τ, they will be integrated into a common track. Otherwise, a new track is created. In our experiments we set $\tau = 200mm$.

Due to noisy detections, occlusions or motion blur, some joints or even full poses might be missing in some frames or noisy. We fill in missing joints by temporal averaging and we smooth each joint trajectory by a Gaussian kernel with standard deviation σ. This simple approach significantly boosts the performance of our model as we will show in Section 4.

	[7]*	[23]*	[34]*	[3]	[4]	[13]	Ours	Ours+
ua	.60	.89	1.0	.68	.98	.97	.99	1.0
la	.35	.68	1.0	.56	.72	.95	.99	1.0
ul	1.0	1.0	1.0	.78	.99	1.0	.98	.99
ll	.90	.99	1.0	.70	.92	.98	.93	.997
avg	.71	.89	1.0	.68	.90	.98	.97	**.997**

Table 1. Quantitative comparison of methods for single human 3D pose estimation from multiple views on the KTH Football II [23] dataset. The numbers are the PCP score in 3D with $\alpha = 0.5$. Methods annotated with * can only estimate single human poses, discretize the state space and rely on being provided with a tight 3D bounding box centered at the true 3D location of the person. *Ours+* and *Ours* describe our method with and without track smoothing (Section 3.2). *ul* and *la* show the scores for upper and lower arm, respectively, while *ul* and *ll* represent upper and lower legs.

4   Experiments

We evaluate our approach on two human pose estimation tasks, single person 3D pose estimation and multi-person 3D pose estimation, and compare it to state-of-the-art methods. Percentage of correct parts (PCP) in 3D as described in [7] is used for evaluation. We evaluate on the limbs only as annotated head poses vary significantly throughout various datasets. In all experiments, the order in which the cameras are processed is given by the dataset. We then evaluate the tracking performance. The source code is made publicly available [1].

4.1   Single Person 3D Pose Estimation

Naturally, first works on 3D human pose estimation from multiple views cover only single humans. Typical methods [7, 23, 34] find a solution over the complete discretized state space which is intractable for multiple persons. However, we report their results for completeness. All models were evaluated on the complete first sequence of the second player of the KTH Football II [23] dataset. Our results are reported in Table 1. Our model outperforms all other multi-person approaches and gets close to the state-of-the-art for single human pose estimation [34] which makes strong assumptions and is much more constrained. Our model has the lowest accuracy for lower legs (*ll*) which experience strong deformation and high movement speed. This can be mostly attributed to the 2D pose estimation framework which confuses left and right under motion blur, as can be seen in Figure 7. When smoothing the trajectory (Section 3.2) this kind of errors can be reduced.

---

[1] https://github.com/jutanke/mv3dpose

Campus dataset $(\alpha = 0.5)$

Actor	[3]			[4]			[13]			Ours			Ours$^+$		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
ua	.83	.90	.78	.97	.97	.90	.97	.94	93	.86	.97	.91	.99	.98	.98
la	.78	.40	.62	.86	.43	.75	.87	.79	70	.74	.64	.68	.91	.70	.92
ul	.86	.74	.83	.93	.75	.92	.94	.99	88	1.0	.99	.99	1.0	.98	1.0
ll	.91	.89	.70	.97	.89	.76	.97	.95	81	1.0	.98	.99	1.0	.98	.99
avg	.85	.73	.73	.93	.76	.83	.94	.93	.85	.90	.90	.89	.98	.91	.98
avg*	.77			.84			.91			.90			**.96**		

Shelf dataset $(\alpha = 0.5)$

Actor	[3]			[4]			[13]			Ours			Ours$^+$		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
ua	.72	.80	.91	.82	.83	.93	.93	.78	.94	.99	.93	.97	.1.0	.97	.97
la	.61	.44	.89	.82	.83	.93	.83	.33	.90	.97	.57	.95	.99	.64	.96
ul	.37	.46	.46	.43	.50	.57	.96	.95	.97	.998	1.0	1.0	1.0	1.0	1.0
ll	.71	.72	.95	.86	.79	.97	.97	.93	.96	.998	.99	1.0	1.0	1.0	1.0
avg	.60	.61	.80	.73	.74	.85	.92	.75	.94	.99	.87	.98	.998	.90	.98
avg*	.67			.77			.87			.95			**.96**		

Table 2. Quantitative comparison of multi-person 3D pose estimation from multiple views on the evaluation frames of the annotated Campus [16, 3] and Shelf dataset [3]. The numbers are the PCP score in 3D with $\alpha = 0.5$. *Ours$^+$* and *Ours* describe our method with and without track smoothing (Section 3.2). We show results for each of the three actors separately as well as averaged for each method (*average**).

4.2   Multi-Person 3D Pose Estimation

To evaluate our model on multi-person 3D pose estimation, we utilize the Campus [16, 3], Shelf [3], CMU Panoptic [22] and UMPM [1] dataset. The difficulty of the Campus dataset lies in its low resolution (360×288 pixel) which makes accurate joint detection hard. Furthermore, small errors in triangulation or detection will result in large PCP errors as the final score is calculated on the 3D joint locations. As in previous works [3, 4] we utilize frames $350 - 470$ and frames $650 - 750$ of the Campus dataset and frames $300 - 600$ for the Shelf dataset. Clutter and humans occluding each others make the Shelf dataset challenging. Nevertheless, our model achieves state-of-the-art results on both datasets by a large margin which can be seen in Table 2. Table 3 reports quantitative results on video *p2_chair_2* of the UMPM [1] benchmark. A sample frame from this benchmark can be seen in Figure 4. As the background is homogeneous and the human actors maintain a considerable distance to each other the results of our method are quite strong.

	Ours$^+$	
Actor	1	2
ua	.997	.98
la	.98	.996
ul	1.0	1.0
ll	.99	.997
avg	0.99	0.99

Table 3. Quantitative comparison of multi-person 3D pose estimation from multiple views on *p2_chair_2* of the UMPM benchmark [1].

	Ours	Ours$^+$
160422_ultimatum1 [22]	.89	.89
160224_haggling1 [22]	.92	.92
160906_pizza1 [22]	.92	.93

Table 4. Quantitative evaluation of multi-person 3D pose tracking on the CMU Panoptic dataset [22] using the MOTA [6] score. *Ours$^+$* and *Ours* describe our method with and without track smoothing (Section 3.2).

4.3 Tracking

For evaluating the tracking accuracy, we utilize the MOTA [6] score which provides a scalar value for the rate of false positives, false negatives, and identity switches of a track. Our model is evaluated on the CMU Panoptic dataset [22] which provides multiple interacting people in close proximity. We use videos *160224_haggling1* with three persons, *160422_ultimatum1* with up to seven person, and *160906_pizza1* with six persons. For the videos *160422_ultimatum1* we use frames 300 to 3758, for *160906_pizza1* we use frames 1000 to 4458 and for *160224_haggling1* we use frames 4209 to 5315 and 6440 to 8200. The first five HD cameras are used. Our results are reported in Table 4 which shows that our approach yields strong tracking capabilities.

4.4 Effects of Smoothing

As can be seen in Table 1 and Table 2 the effects of smoothing can be significant, especially when detection and calibration are noisy as is the case with the Campus and the KTH Football II dataset. In both datasets 2D human pose detection is challenging due to low resolution (Campus) or strong motion blur (KTH Football II). Datasets with higher resolution and less motion blur like the Shelf dataset do not suffer from this problems as much and as such do not benefit the same way from track smoothing. However, a small gain can still be noted as smoothing also fills in joint detections that could not be triangulated. Figure 5 explores different σ values for smoothing on the KTH Football II, Campus, and
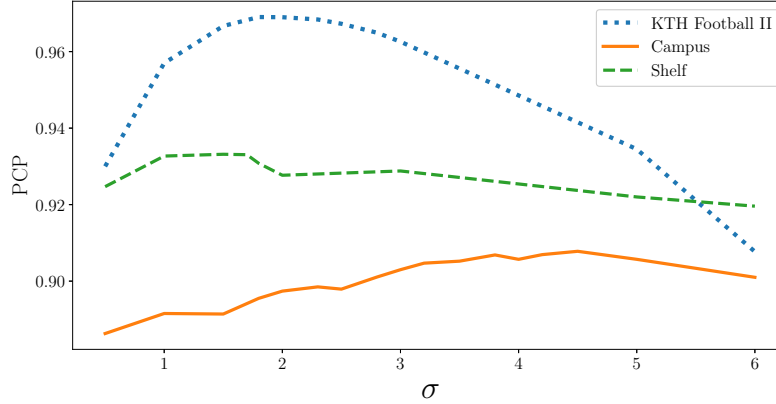
Fig. 5. PCP score for different smoothing values σ for tracking on KTH Football II, Campus, and Shelf. If σ is too small, the smoothing has little effect and coincides with the un-smoothed results. When the joint trajectories are smoothed too much, the PCP score drops as well as the trajectories do not follow the original path anymore. (Larger PCP scores are better)

Shelf dataset. It can be seen that smoothing improves the performance regardless of the dataset but that too much smoothing obviously reduces the accuracy. We chose $\sigma = 2$ for all our experiments except for the Campus dataset where we set $\sigma = 4.2$. The reason for the higher value of σ for the Campus dataset is due to the very low resolution of the images compared to the other datasets, which increases the noise of the estimated 3D joint position by triangulation.

4.5   Effects of camera order

So far we used the given camera order for each dataset, but the order in which views are greedily matched matters and different results might happen with different orderings. To investigate the impact of the camera order, we evaluated our approach using all 120 permutations of the 5 cameras of the Shelf dataset. The results shown in Figure 6 show that the approach is very robust to the order of the camera views.

4.6   Early Commitment

A failure case happens due to the early commitment of our algorithm with regards to the 2D pose estimation, as can be seen in Figure 7. When the pose estimation is unsure about a pose, it still fully commits to its output and disregards uncertainty. This problem occurs due to motion blur as the network has difficulties to decide between left and right in this case. As our pose estimation
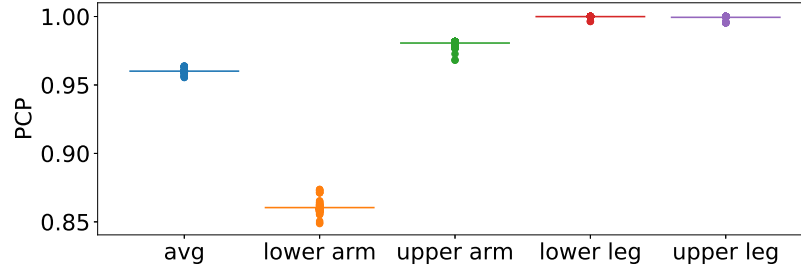
Fig. 6. PCP score averaged over all subjects for all 120 camera permutations of the Shelf dataset. The vertical line represents the mean value over all permutations while the dots represent each camera permutation.
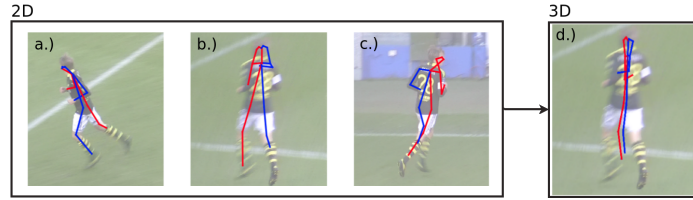

Fig. 7. Issues with early commitment. As we utilize the 2D pose estimations directly, our method suffers when the predictions yield poor results. In this example the pose estimation model correctly estimates (a) and (c) but confuses left and right on (b) due to motion blur. The resulting 3D pose estimation (d) collapses into the centre of the person. The red limbs represent the right body side while blue limbs represent the left body side.

model has mostly seen forward-facing persons it will be more inclined towards predicting a forward-facing person in case of uncertainty. When left and right of a 2D prediction are incorrectly flipped in at least one of the views, the merged 3D prediction will collapse to the vertical line of the person resulting in a poor 3D pose estimation.

5 Conclusion

In this work we presented a simple baseline approach for 3D human pose estimation and tracking from multiple calibrated cameras and evaluate it extensively on several 3D multi-camera datasets. Our approach achieves state-of-the-art results in multi-person 3D pose estimation while remaining sufficiently efficient for fast processing. Due to the models simplicity some common failure cases can be noted which can be build upon in future work. For example, confidence maps

provided by the 2D pose estimation model could be utilized to prevent left-right flips. Our approach may serve as a baseline for future work.

Acknowledgement

References

1. Aa, N.v.d., Luo, X., Giezeman, G., Tan, R., Veltkamp, R.: Utrecht Multi-Person Motion (UMPM) benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: Workshop on Human Interaction in Computer Vision (2011)
2. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view Pictorial Structures for 3D Human Pose Estimation. In: British Machine Vision Conference (2013)
3. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Conference on Computer Vision and Pattern Recognition (2014)
4. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures revisited: Multiple human pose estimation. Transactions on Pattern Analysis and Machine Intelligence (2016)
5. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A study of parts-based object class detection using complete graphs. International Journal of Computer Vision (2010)
6. Bernardin, K., Elbs, A., Stiefelhagen, R.: Multiple object tracking performance metrics and evaluation in a smart room environment. In: Workshop on Visual Surveillance (2006)
7. Burenius, M., Sullivan, J., Carlsson, S.: 3D pictorial structures for multiple view articulated pose estimation. In: Conference on Computer Vision and Pattern Recognition (2013)
8. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: Conference on Computer Vision and Pattern Recognition (2017)
9. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Conference on Computer Vision and Pattern Recognition (2018)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition (2005)
11. Doering, A., Iqbal, U., Gall, J.: JointFlow: Temporal Flow Fields for Multi Person Tracking. British Machine Vision Conference (2018)
12. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: Conference on Computer Vision and Pattern Recognition (2015)
13. Ershadi-Nasab, S., Noury, E., Kasaei, S., Sanaei, E.: Multiple human 3d pose estimation from multiview images. Multimedia Tools and Applications (2018)

14. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision (2005)
15. Fieraru, M., Khoreva, A., Pishchulin, L., Schiele, B.: Learning to refine human pose estimation. In: Conference on Computer Vision and Pattern Recognition Workshops (2018)
16. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. Pattern Analysis and Machine Intelligence (2007)
17. Guo, H., Tang, T., Luo, G., Chen, R., Lu, Y., Wen, L.: Multi-Domain Pose Network for Multi-Person Pose Estimation and Tracking. In: European Conference on Computer Vision (2018)
18. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision (2016)
19. Iqbal, U., Doering, A., Yasin, H., Krüger, B., Weber, A., Gall, J.: A dual-source approach for 3D human pose estimation from single images. Computer Vision and Image Understanding (2018)
20. Iqbal, U., Milan, A., Gall, J.: PoseTrack: Joint Multi-Person Pose Estimation and Tracking. In: Conference on Computer Vision and Pattern Recognition (2017)
21. Iqbal, U., Molchanov, P., Breuel Jürgen Gall, T., Kautz, J.: Hand pose estimation via latent 2.5D heatmap regression. In: Proceedings of the European Conference on Computer Vision (2018)
22. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic Studio: A Massively Multiview System for Social Motion Capture. In: International Conference on Computer Vision (2015)
23. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: British Machine Vision Conference (2013)
24. Kocabas, M., Karagoz, S., Akbas, E.: MultiPoseNet: Fast multi-person pose estimation using pose residual network. In: European Conference on Computer Vision (2018)
25. Kostrikov, I., Gall, J.: Depth Sweep Regression Forests for Estimating 3D Human Pose from Images. In: British Machine Vision Conference (2014)
26. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: Conference on Computer Vision and Pattern Recognition1 (2011)
27. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: International Conference on Computer Vision (2017)
28. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: International Conference on 3D Vision (2017)
29. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular RGB. In: International Conference on 3D Vision (2018)
30. Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics (1957)
31. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems (2017)
32. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (2016)

33. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Conference on Computer Vision and Pattern Recognition (2017)
34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting Multiple Views for Marker-less 3D Human Pose Annotations. In: Conference on Computer Vision and Pattern Recognition (2017)
35. Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F., Müller, E., Salzmann, M., Fua, P.: Learning Monocular 3D Human Pose Estimation from Multi-view Images. In: Conference on Computer Vision and Pattern Recognition (2018)
36. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. Transactions on Pattern Analysis and Machine Intelligence (2019)
37. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: International Conference on Computer Vision (2011)
38. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. Conference on Computer Vision and Pattern Recognition (2017)
39. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (2018)
40. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: International Conference on Computer Vision (2015)