3D Pictorial Structures for Human Pose Estimation with Supervoxels

Alexander Schick
Fraunhofer IOSB
alexander.schick@iosb.fraunhofer.de

Rainer Stiefelhagen
Karlsruhe Institute of Technology
rainer.stiefelhagen@kit.edu

Abstract

Pictorial structures provide a powerful framework for human pose estimation, in particular in the domain of 2D data. However, solving pictorial structures directly in 3D drastically increases its complexity and it quickly exceeds tractable dimensions. In this paper, we propose a discretization-by-segmentation approach by applying supervoxels to 3D pictorial structures which significantly reduces the search space. The proposed 3D pictorial structures approach achieves 3D errors of 115 *mm and* 135 *mm on the HumanEva-I and UMPM datasets and PCP scores of* 78% *and* 75%*, respectively. Due to the search space reduction, the overall pose estimation runtime is below* 100 *ms which is up to four orders of magnitude faster than comparable 3D pictorial structure approaches. The presented approach is not limited to human pose estimation, but provides a general and efficient solution for 3D pictorial structures.*

1. Introduction

Human pose estimation is the problem of determining all free parameters of a body model. In particular for human pose estimation, the number of parameters is very large and degrees of freedom of 30 and more are quite common. This results in a huge search space with a large number of possible poses. Consequently, each pose estimation algorithm requires an efficient strategy to cope with this large search space. In this work, we will discuss such strategies and show how supervoxels can be used to reduce the search space to tractable dimensions.

One particular framework for pose estimation are pictorial structures which proved to be very useful for 2D pose estimation. In 3D, however, pictorial structures are more difficult to apply. The additional dimension affects all free parameters and significantly increases its complexity and search space. In this work, we apply supervoxels to pictorial structures to allow for 3D pose estimation in under 100 ms. To the best of our knowledge, this makes it the most efficient 3D pictorial structures approach.

2. Related Work

In this section, we discuss related pose estimation approaches with a focus on strategies to cope with the large search space as well as pictorial structures. In addition, we explain the concept of supervoxels and their advantages.

2.1. Pose Estimation

Human pose estimation is a very active field of computer vision as it allows for many commercial and scientific applications, for example surveillance, motion analysis, and human-computer interaction, to name just a few. There exist several extensive reviews, for example Moeslund and Granum [19], Moeslund *et al.* [20], Poppe [23], and Holte *et al.* [13]. Here, we concentrate on 3D human pose estimation with a focus on strategies to efficiently address the large search space. In particular, we would like to discuss three specific strategies: tracking, detection, and discretization.

Tracking approaches use the continuity of movements to reduce the search space. By estimating where a moving body part is likely to be in the next observation, the search can concentrate around this location. Various tracking algorithms exist, for example Kalman filters [14] or particle filters [6, 7]. The main drawback of tracking is that it requires initialization. In addition, in case of tracking failure, a recovery mechanism is necessary for reinitialization.

Detection-based approaches learn part detectors, *e.g.,* for the head or hands, and apply them to the observation. Then, the search space for body parts can be reduced to areas where the respective detectors determined a high likelihood. Detection-based approaches showed excellent results [28, 31]. However, there are two main drawbacks. First, they require a large set of training data. For example, Shotton *et al.* [28] used approximately 300,000 labeled training images. Second, the training data introduces a bias. In case of [28], a rotation of the camera is sufficient for the part detectors to fail.

The last strategy that we discuss is discretization. Here, the search space is discretized into fixed intervals. This could, for example, be a grid with respect to positions or fixed angle intervals with respect to joint angles. Discretization is very effective [5], but also suboptimal as it ignores

the underlying data. This leads to two main drawbacks. First, discretization intervals are sampled irrespective of the data and can therefore be outside the currently observed body volume. Second, these intervals are fixed and do not adjust to the data. This means that the true position of a body part can be exactly between two search intervals.

In this work, we present an approach that improves over current discretization solutions. We propose to use discretization-by-segmentation as preprocessing. By segmenting the voxels into supervoxels, the number of input elements (*i.e.,* voxels) is significantly reduced while still conforming to the observed data.

2.2. Pictorial Structures

Pictorial structures have been proposed by Fischler and Elschlager [11]. A pictorial structure is a simplified way to describe an object. It consists of two elements: atomic object parts and connections between these parts. An efficient solution for pictorial structures was proposed by Felzenszwalb and Huttenlocher [9]. They showed how pictorial structures can be computed efficiently with dynamic programming if the representation has no cycles and showed applications to 2D face detection and human pose estimation.

Pictorial structures are well-suited for 2D human pose estimation ([2, 4, 10]). For 3D human pose estimation, however, the search space quickly exceeds tractable dimensions [5]. One of the few approaches that solves pictorial structures in 3D was proposed by Burenius *et al.* [5]. They showed how the 3D search space can be reduced by discretization and they achieved runtimes between 1 s and 69 min for each frame. In this paper, we show how supervoxels provide a better discretization for 3D pictorial structures with runtimes well below 100 ms.

2.3. Supervoxels

Supervoxels are the continuation of superpixels in 3D. The concept of superpixel segmentation was introduced by Ren and Malik [34]. A superpixel describes a connected region of similar pixels that are close together. Superpixels have become quite popular and there exist a large variety of superpixel algorithms, for example [1, 16, 18, 27, 33]. The reason is that superpixels effectively cluster similar pixels while still conforming to object boundaries. Therefore, they are well suited to be used as atomic primitives in applications instead of pixels. As one superpixel usually consists of hundreds of pixels, this leads to a significant reduction of the number of input elements, often in the range of several orders of magnitude.

Supervoxels apply the concept of superpixels to 3D data. 3D data can either be created by stacking images, for example video frames [33, 35] or medical scans [3, 17], or by using real 3D data [22]. In this work, we are mostly interested in volumetric 3D data from multi-view voxel carving. However, it can also be applied to non-volumetric 3D point clouds, *e.g.,* based on stereo or RGB+D sensors.

One of the few supervoxel approaches for 3D data was proposed by Papon *et al.* [22]. In this work, we use an algorithm based on [25], but extended to 3D data, that provides similar results. However, it is very important to note that the pose estimation algorithm presented here does not depend on a specific supervoxel algorithm, as it does not make assumptions about the segmentation. This also implies that our approach directly benefits from future supervoxel algorithms with better segmentation performance.

In summary, supervoxels cluster similar voxels that are close together. Therefore, they adjust to the data and provide a data-dependent discretization. Similar to superpixels, the number of data elements, the voxels, is significantly reduced. As we will show in the remainder of this paper, this discretization-by-segmentation as well as how we apply supervoxels to pose estimation leads to a low computational complexity and low runtime.

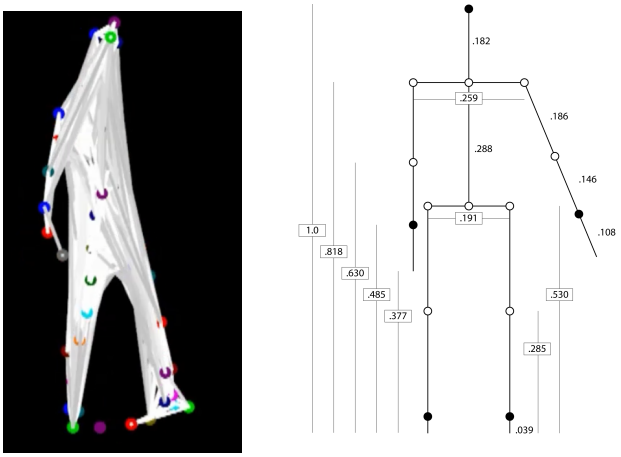
3. Human Pose Estimation with Supervoxels

This section explains how we apply supervoxels to the pictorial structures framework. The key idea of the proposed algorithm, partially motivated by [21], is to restrict joints of the human body model to supervoxel centers. This assumption significantly reduces the search space because the number of supervoxels is much smaller than the number of voxels or 3D points. And because supervoxels conform to the observed 3D data, they still provide an adequate representation. As we will show in Section 5, this leads to a reduction of several orders of magnitude.

3.1. Improved Discretization with Supervoxels

Discretization typically partitions the search space and uses fixed step sizes. This is data agnostic and, therefore, suboptimal. For example, with 10° joint angle intervals, there would be 36 search intervals independent of the actual data. In the worst case, positions are sampled where actually no supporting data exists. A discretization based on the data would be much better.

Supervoxels provide such an improved discretization. First, supervoxels cluster similar voxels that are close together. Therefore, they are likely to belong only to one body part which increases accuracy. Second, supervoxels exist only where actual observations exist. Therefore, every supervoxel is a valid candidate and it cannot occur that positions outside the human body are chosen. Restricting joints to supervoxels centers has an additional advantage: the number of body part candidates is also reduced as will be explained in the context of the supervoxel graph.



(a) Weighted supervoxel graph

(b) Anthropometric ratios

Figure 1. Supervoxel graph and body model. (a) shows an example of the supervoxel graph with lighter connections having a higher weight. (b) shows the anthropometric ratios used in this work (based on [8] cited by [12]).

3.2. Supervoxel Graph

By restricting joints to supervoxel centers, it then follows that potential candidates for body parts are restricted to connections between supervoxel centers. These connections can be represented by the *supervoxel graph* $G = (\mathcal{S}, E)$ with supervoxels $s \in \mathcal{S}$ and edges $e \in E$. The supervoxel graph is different from an adjacency graph [22] that only connects neighboring supervoxels, as it includes all possible connections betweens supervoxels. An example of a supervoxel graph is shown in Figure 1a.

In case of volumetric voxel data, as is the case in this paper, the connections can be weighted by the fraction they lie within the segmented voxel volume. This actually resembles limbs that are also within the body and is useful to reduce the overall number of limb candidates as connections outside the segmented volume can be filtered out. In case of 3D point clouds, a similar behavior can be achieved by checking if 3D points are sufficiently close to a supervoxel connection. In both cases, this can be efficiently evaluated by using kd-trees or by a direct lookup in the voxel grid cells.

This concludes the preprocessing that provides a supervoxel segmentation and its corresponding supervoxel graph. The next section will show how these can be applied to pictorial structures.

4. 3D Pictorial Structures with Supervoxels

Pictorial structures describe an object with its atomic parts and connections between them [9, 11]. In this work, we model the human body with twelve body parts, ten connecting joints, and five end joints. The body model is shown in Figure 1b.

Following Felzenszwalb and Huttenlocher [9], pose estimation can follow a Bayesian formulation and therefore be expressed as an energy minimization problem. Let θ be the body model and let L be a specific configuration of the N body parts $l_1, l_2, \cdots, l_N$. Let the unary term m_i, $1 \leq i \leq N$, be the energy of body part i and let the binary term d_{ij} give the energy of the connection between two body parts. Then, the optimal configuration L^* minimizes the following energy function:

$$L^* = \underset{L}{\operatorname{argmin}}\Big(\sum_{i}^{N} m_i(l_i) + \sum_{(i,j)\in\theta} d_{ij}(l_i, l_j)\Big). \quad (1)$$

Given this energy function, we will now show how the unary and binary energy terms are computed. The unary term m_i, or data term, measures how well a body part candidate matches the expected appearance. Because we work in the domain of 3D data, we use anthropometric ratios from [8] cited by [12] (Figure 1b) to compute expected limb lengths given the overall body height. The expected lengths are then compared to connections of the supervoxel graph. Let $||l_i||$ be the length of body part candidate l_i and let $||\hat{l}_i||$ be the expected length of body part i. Then, the unary energy term is given by

$$m_i(l_i) = \frac{\big|\,||l_i|| - ||\hat{l}_i||\,\big|}{||\hat{l}_i||}. \quad (2)$$

The binary energy term consists of two separate terms that are combined. The first term enforces that two body parts i and j can only be connected if and only if they share the same supervoxel as connecting joint c_{ij}. This term actually reduces the complexity as not all pairs of limbs must be compared, but only limbs that share a connecting supervoxel:

$$d^s_{ij}(l_i, l_j) = \begin{cases} 0 & \text{if } c_{ij}(l_j) = c_{ij}(l_i) \\ \infty & \text{otherwise} \end{cases}. \quad (3)$$

The second binary energy term is more specific to the volumetric voxel data used in this work. As volumeless skeletons can be fit very tightly into volumes, we prefer extended limbs. This heuristic pose prior works well for volumes as it ensures that the given volume is completely filled. Let g_{ij} give the distance between the end joints of two connected limbs i and j (*e.g.,* shoulder and hand in case

of upper and lower arm) and let $\hat{g}_{ij}$ give the expected distance if they are fully extended. The second binary energy term is then given by

$$d^e_{ij}(l_i, l_j) = \frac{1}{\frac{g_{ij}(l_i,l_j)}{\hat{g}_{ij}(l_i,l_j)}} = \frac{\hat{g}_{ij}(l_i, l_j)}{g_{ij}(l_i, l_j)}. \quad (4)$$

As stated above, this energy term is specific for volumetric data. Depending on the task at hand, there are two alternatives. First, the pose prior can be learned from training data. We investigated this solution but found no improvement for volumetric data. Second, by introducing part detectors, the most likely joint positions can be additionally modeled in the unary energy term. This would provide anchors for the joints, thus alleviating the need for this pose prior.

4.1. Efficient Estimation of Human Poses

Felzenszwalb and Huttenlocher [9] showed that pictorial structures can be solved efficiently if the object is represented by a tree. This is the case for the human body model (Figure 1b). One way to efficiently estimate the pose is through dynamic programming. Here, we use the min-sum algorithm as was also done by Burenius *et al.* [5].

Algorithm 1 Min-sum pose estimation with supervoxels

Input: supervoxel graph $G = (\mathcal{S}, E)$
Input: number of pose candidates P
Output: pose configurations $L^*_1, \cdots, L^*_P$
Define $p(i)$: return parent node of body part i
// Initialization
$\forall l \in E \; \forall n \in \{1, \cdots, N\}$: initialize part scores $m_n(l_n)$
// Passing messages upwards
for $n := N$ *to* 2 **do**
  for $l_{p(n)} \in E$ **do**
    $\overline{m} := \min_{l_n}(d^s_{n,p(a)}(l_n, l_{p(n)}) + m_n(l_n))$
    $m_{p(n)} := m_{p(n)} + \overline{m}$
  end for
end for
// Passing messages downwards
for $i := 1$ *to* P **do**
  $l^*_1 := \operatorname{argmin}_{l_1}(m_1(l_1))$
  $L^*_i(1) := l^*_1$
  for $n := 2$ *to* N **do**
    $l^*_n := \operatorname{argmin}_{l_n}(d^e_n(l_n, L^*) + m_n(l_n))$
    $L^*_i(n) := l^*_n$
  end for
end for

After initializing each connection of the supervoxel graph with the unary energies of each body part, the final pose is computed through two message passing phases. First, the unary terms are propagated upwards the kinematic chain while also including the respective connections between body parts. Then, the best torso candidate is selected and messages are passed downwards to select the best body parts given the selected torso. Algorithm 1 shows pseudo code for this procedure.

As explained above, we use a second energy term to prefer connections that fill the available volume. This term is only used during the second part of Algorithm 1. Therefore, the min-sum algorithm becomes a greedy algorithm. For best results, we sample additional poses by choosing the n best torso candidates at the beginning of the second part. As we will discuss in Section 5, pose candidates can be efficiently evaluated and the best solution is found with only a small set of samples.

This concludes the introduction of the 3D pictorial structures approach with supervoxels. Additional information can be found in [24]. The next section shows the evaluation and a discussion of the parameters.

5. Evaluation

This section presents the evaluation of the proposed 3D pictorial structures algorithm with supervoxels. First, the datasets and error metrics are introduced. Then, variations of the parameters are shown followed by a final evaluation with the best parameters including comparisons to other work. The section is concluded with a complexity and runtime evaluation.

For all experiments, an Intel Pentium Intel(R) Core(TM) i7-3770 CPU with 3.40 GHz was used. GPU code was implemented with NVIDIA Cuda and executed on a NVIDIA GeForce GTX 660 Ti.

5.1. Datasets and Metrics

We used two datasets for evaluation. The HumanEva-I (HE-I) dataset introduced by Sigal and Black [30] and Sigal *et al.* [29] and the Utrecht Multi-Person Motion (UMPM) benchmark introduced by Van der Aa *et al.* [32]. Both datasets provide multiple views of calibrated cameras that are required for voxel carving. Ground truth was measured with Vicon systems. While HE-I shows only action sequences, UMPM also includes interactions with large objects like chairs and tables. We used all five action sequences of actors S1 and S2 of the HE-I dataset and all five sequences showing only one actor of the UMPM dataset. In total, more than $12,500$ frames were used.

For evaluation, we applied two error metrics: The first metric is the relative 3D joint localization error (Sigal *et al.* [29], Equation 6) that measures the average distance of estimated joint positions to the ground truth in millimeters. The second metric, percentage of correct parts (PCP) [10], measures the fraction of correctly estimated parts over the

(a) S1 ThrowCatch sequence, voxel size: 2 cm, supervoxel size: 10 cm

(b) Table sequence, voxel size: 2 cm, supervoxel size: 10 cm, synthetic voxels

Figure 2. Qualitative pose estimation results for the HumanEva-I and UMPM datasets for various voxel and supervoxel sizes.

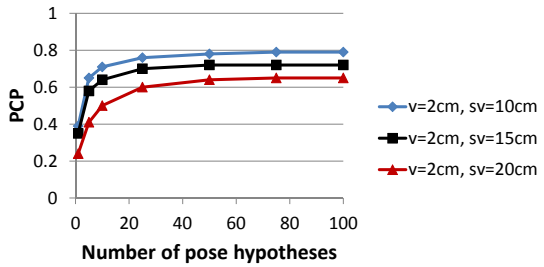
whole sequence. A part is counted as correctly estimated if and only if both joints are within a distance of at most half the limb length. This metric is often used for 2D pictorial structures approaches and we use it here for a better comparison to future work with 3D pictorial structures.

5.2. Parameter Evaluation

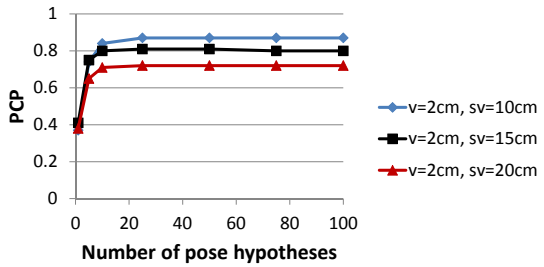
The presented approach has three main stages: voxel carving, supervoxel segmentation, and pose estimation. Here, we focus on the last stage. For voxel carving, we used the algorithm presented by Schick and Stiefelhagen [26] that provides a mechanism to reason with static occlusions which is necessary for the UMPM dataset. For supervoxel segmentation, we adapted the approach presented in [25] to 3D voxel data. The supervoxels were used with a compactness parameter of $\alpha = 1.0$ [25] as the voxels are colorless.

The whole approach requires few parameters and their influence on pose estimation will now be evaluated. The parameters are the number of pose hypotheses (Section 4.1), voxel and supervoxel sizes, and supervoxel graph weight thresholds (Section 3.2).

Figure 3 shows variations of the number of pose hypotheses for different voxel and supervoxel sizes. The results stabilize quickly at approximately 50 hypotheses which shows that the greedy version of the min-sum algorithm requires only few sampled poses for best results. This number of hypotheses is used for the remaining evaluations.



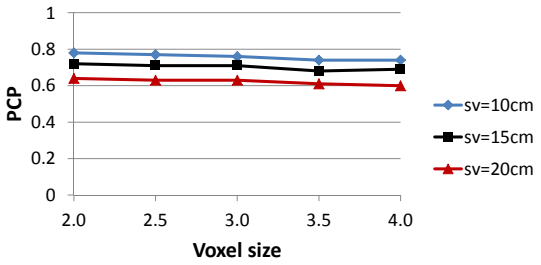
(a) HumanEva-I



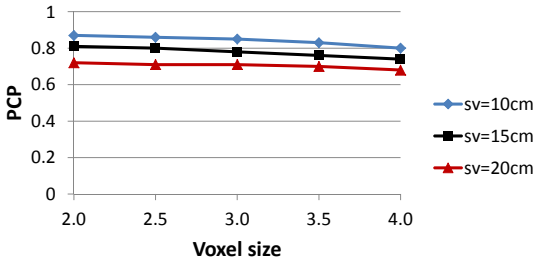
(b) UMPM (Orthosyn sequence)

Figure 3. Evaluation of the number of pose hypotheses for the HEI and UMPM datasets.

Figure 4 shows evaluations of various voxel sizes and three supervoxel sizes. While the voxel size has only a small impact on overall PCP results, the supervoxels are the de-

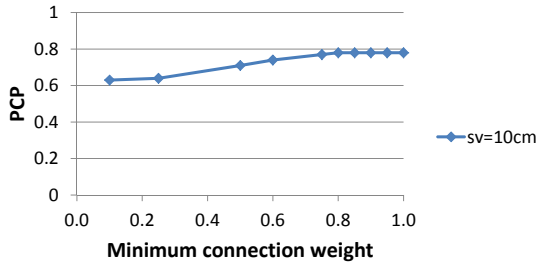


(a) HumanEva-I

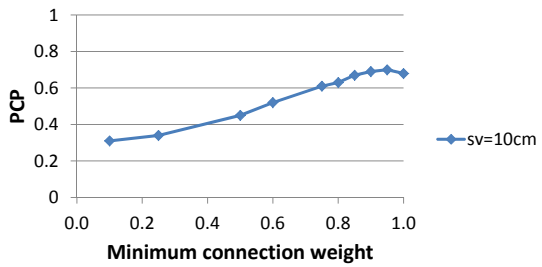


(b) UMPM (Orthosyn sequence)

Figure 4. Evaluation of voxel and supervoxel sizes for the HE-I and UMPM datasets.



(a) PCP scores for HumanEva-I



(b) PCP scores for UMPM (chair sequence)

Figure 5. Evaluation of supervoxel graph connection weights for the HE-I and UMPM datasets.

termining factor. This shows that supervoxels abstract very well from the underlying voxel data. It also implies that future supervoxel algorithms with better performance will directly improve pose estimation without any modification of our proposed algorithm. If not specified otherwise, a voxel size of 2 cm and supervoxel size of 10 cm will be used in subsequent evaluations.

Figure 5 shows results for varying supervoxel graph weights. For HE-I, results are best with a very strict weight of 1.0 which means that there is no tolerance for connections outside the segmented volume. For UMPM, best results are achieved with a connection weight threshold of 0.95. The reason is that there are more voxel carving errors for UMPM due to static occlusions. In the remainder of the evaluation, the thresholds are 1.0 for HE-I and 0.95 for UMPM.

5.3. Evaluation with Optimal Parameters

The final evaluation presents both 3D error and PCP results for three voxel and supervoxel sizes with 50 pose hypotheses and connection weights of 1.0 and 0.95, respectively. First, results on synthetic data are shown. Synthetic data consists of voxels sampled around the ground truth body parts and is used to verify that the approach works correctly given good voxel data. Second, results with voxel carving are presented (as was done in Section 5.2). These include voxel carving errors due to difficult foreground segmentations for the HE-I and static occlusions for the UMPM dataset.

Table 5.3 shows PCP results for synthetic voxel data for

Dataset	Voxel: 2 cm SV: 10 cm	Voxel: 3 cm SV: 15 cm	Voxel: 4 cm SV: 20 cm
HE-I (syn)	**96**	92	86
HE-I (real)	**78**	71	60
UMPM (syn)	**98**	95	89
UMPM (real)	**75**	66	60

Table 1. PCP results for pose estimation with both synthetic as well as voxel carving data on the HumanEva-I and UMPM datasets. The table shows results for three voxel and supervoxel resolutions.

Dataset	Voxel: 2 cm SV: 10 cm $[mm]$	Voxel: 3 cm SV: 15 cm $[mm]$	Voxel: 4 cm SV: 20 cm $[mm]$
HE-I	**115**	139	181
UMPM	**135**	167	185

Table 2. 3D joint localization error for pose estimation based on voxel carving data for the HumanEva-I (HE) and UMPM datasets. The table shows results for three voxel and supervoxel resolutions.

both datasets. Even for rather large voxels and supervoxels, the PCP results are close to 90%. For smaller voxels and supervoxels, the results are close to 100%. Table 5.3 also includes PCP results for both datasets on voxel carving data. While the results are worse due to imperfect voxel carving, they show recognition rates of 78% and 75%.

Table 5.3 shows 3D errors for both datasets over all frames. As expected, the errors increase for larger voxels

and supervoxels. For the smallest resolutions (10 cm supervoxels), results are 115 mm for HE-I and 135 mm for UMPM.

5.4. Comparison to Related Work

A comparison of PCP results is difficult as this metric is mostly used for pictorial structures approaches in 2D. However, PCP results were reported by Burenius *et al.* [5]. Their approach is closest to our algorithm as they also solve pictorial structures in 3D. They reported PCP results of up to 77%, but on a custom dataset. These results, however, are in the same range as our reported results in Section 5.3.

Most 3D approaches evaluate the 3D error. Canton-Ferrer *et al.* [6] proposed an annealed particle filter based on voxel data with a resolution of 2 cm (as in this work) and evaluated on the HE-I dataset. Their approach is computationally very expensive as $3,600$ hypotheses must be tracked simultaneously. They report 3D errors of 121.18 mm.

Amin *et al.* [2] proposed an approach that computes 2D poses with pictorial structures on multiple camera images that are then triangulated in 3D. They achieve among the best 3D errors and report results between 44.7 mm and 62.4 mm. However, their approach requires separate computations of multiple 2D pictorial structures which is computationally expensive.

Kanaujia *et al.* [15] used part detectors similar to Shotton *et al.* [28], but trained them on voxel data. They evaluated their system on various sequences of the HE-I dataset and achieved 3D errors between 71.261 mm and 90.952 mm.

3D errors for the UMPM dataset cannot be compared as there are no reported results for similar approaches available yet.

5.5. Runtime and Complexity Analysis

The evaluation in Table 5.5 shows runtimes for all steps of our approach. The runtimes are reported for the GPU implementations. The pose estimation runtimes are 91 ms for the smallest voxel and supervoxel resolutions and decrease rapidly for larger ones.

In comparison, the approach of Burenius *et al.* [5], which is closest to our approach as it also solves pictorial structures in 3D, requires runtimes between 1 s and 69 min depending on the discretization parameters. This means that our approach is up to four orders of magnitude faster. The other approaches discussed above did not report any runtimes for comparison.

The complexity of the proposed 3D pictorial structures algorithm depends on the number of supervoxels. Let $N = |\mathcal{S}|$ be the number of supervoxels. Each connection between supervoxels is a potential body part candidate leading to N^2 body part candidates in the worst case.

Algorithm part	SV size: 10 cm $[ms]$	SV size: 15 cm $[ms]$
Voxel carving	18	6
Supervoxels	12	10
Supervoxel graph	1	< 1
Pose estimation	91	14
Total	122	30

Table 3. Runtimes for all steps of the presented pose estimation system for the GPU implementation.

The min-sum algorithm consists of an initialization and two phases, up and down. The time complexity for initialization is in $\mathcal{O}(N^2)$ as each supervoxel graph connection must be initialized. The up phase is the computationally most complex part as each pair of connections must be combined. However, this does not lead to a complexity in $\mathcal{O}(N^4)$ but only $\mathcal{O}(N^3)$ because both connections must share the same connecting supervoxel. The down phase is the least expensive part after the torso has been selected. Then, each selected body part is determined by its end joint as the connecting joint has already been selected, thus leading to a complexity in $\mathcal{O}(N)$. This is also the reason why sampling multiple poses is computationally efficient. In conclusion, the time complexity is determined by the up phase and is in $\mathcal{O}(N^3)$. The space complexity is in the number of body parts and, therefore, in $\mathcal{O}(N^2)$.

6. Conclusion

This concludes our presentation of 3D pictorial structures for human pose estimation with supervoxels. We showed how supervoxels can be applied to pictorial structures to efficiently estimate the human pose. An evaluation on two datasets showed comparable performance in terms of 3D joint localization error and PCP results. The computational efficiency exceeds comparable approaches by up to four orders of magnitude. The presented approach is not limited to human pose estimation, but provides a general and efficient solution for 3D pictorial structures.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–82, 2012.

[2] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view Pictorial Structures for 3D Human Pose Estimation. In *British Machine Vision Conference*, 2013.

[3] B. Andres, U. Köthe, M. Helmstaedter, W. Denk, and F. A. Hamprecht. Segmentation of SBFSEM Volume Data of Neural Tissue by Hierarchical Classification. In *DAGM Symposium on Pattern Recognition*, pages 142–152, 2008.

[4] M. Andriluka, S. Roth, and B. Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In *Computer Vision and Pattern Recognition*, pages 1014–1021, 2009.

[5] M. Burenius, J. Sullivan, and S. Carlsson. 3D Pictorial Structures for Multiple View Articulated Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 3618–3625, 2013.

[6] C. Canton-Ferrer, J. Casas, and M. Pardas. Voxel based annealed particle filtering for markerless 3D articulated motion capture. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pages 1–4, 2009.

[7] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition*, volume 2, pages 126–133, 2000.

[8] R. Drillis and R. Contini. Body segment parameters. Technical Report No. 1166.03. New York University, School of Engineering and Science. Technical report, 1966.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[10] V. Ferrari, M. Marín-Jimenéz, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[11] M. A. Fischler and R. A. Elschlager. The Representation and Matching of Pictorial Structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[12] R. C. Fromuth and M. B. Parkinson. Predicting 5th and 95th percentile anthropometric segment lengths from population stature. In *ASME International Design Engineering Technical Conferences. DETC2008-50091.*, 2008.

[13] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):538–552, 2012.

[14] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1453–1459, 2000.

[15] A. Kanaujia, N. Kittens, and N. Ramanathan. Part Segmentation of Visual Hull for 3D Human Pose Estimation. In *Computer Vision and Pattern Recognition Workshops*, pages 542–549, 2013.

[16] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *Computer Vision and Pattern Recognition*, pages 2097–2104, 2011.

[17] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Transactions on Medical Imaging*, 31(2):474–86, 2012.

[18] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[19] T. B. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3):231—-268, 2001.

[20] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[21] G. Mori. Guiding model search using segmentation. In *International Conference on Computer Vision*, pages 1417–1423 Vol. 2, 2005.

[22] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter. Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds. In *Computer Vision and Pattern Recognition*, pages 2027–2034, 2013.

[23] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1):4–18, 2007.

[24] A. Schick. *Human Pose Estimation with Supervoxels*. PhD thesis, Karlsruhe Institute of Technology. Department of Informatics, May 2014.

[25] A. Schick, M. Fischer, and R. Stiefelhagen. Measuring and evaluating the compactness of superpixels. In *International Conference on Pattern Recognition*, pages 930–934, 2012.

[26] A. Schick and R. Stiefelhagen. Real-Time GPU-Based Voxel Carving with Systematic Occlusion Handling. In *DAGM Symposium on Pattern Recognition*, pages 372–381, 2009.

[27] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[28] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[29] 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. *International Journal of Computer Vision*, 87(1-2):4–27, 2010.

[30] L. Sigal and M. J. Black. HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion. In *Technical Report CS-06-08, Brown University*, number September, 2006.

[31] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Computer Vision and Pattern Recognition*, pages 103–110, 2012.

[32] N. van der Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp. UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *International Conference on Computer Vision Workshops*, pages 1264–1269, 2011.

[33] O. Veksler, Y. Boykov, and P. Mehrani. Superpixels and Supervoxels in an Energy Optimization Framework. In *European Conference on Computer Vision*, pages 211–224, 2010.

[34] Xiaofeng Ren and J. Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision*, pages 10–17, 2003.

[35] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *Computer Vision and Pattern Recognition*, pages 1202–1209, 2012.