Estimating Human 3D Pose from Time-of-Flight Images Based on Geodesic Distances and Optical Flow

Loren Arthur Schwarz Artashes Mkhitaryan Diana Mateus Nassir Navab

Chair for Computer Aided Medical Procedures (CAMP), Technische Universität München, 85748 Garching, Germany
{schwarz,mkhitary,mateus,navab}@cs.tum.edu http://campar.cs.tum.edu

Abstract— **In this paper, we present a method for human full-body pose estimation from Time-of-Flight (ToF) camera images. Our approach consists of robustly detecting anatomical landmarks in the 3D data and fitting a skeleton body model using constrained inverse kinematics. Instead of relying on appearance-based features for interest point detection that can vary strongly with illumination and pose changes, we build upon a graph-based representation of the ToF depth data that allows us to measure geodesic distances between body parts. As these distances do not change with body movement, we are able to localize anatomical landmarks independent of pose. For differentiation of body parts that occlude each other, we employ motion information, obtained from the optical flow between subsequent ToF intensity images. We provide a qualitative and quantitative evaluation of our pose tracking method on ToF sequences containing movements of varying complexity.**

I. Introduction

Human gestures are a natural means of communication and allow conveying complex information. Using gestures for interaction with computer-assisted systems can be of great benefit, particularly in scenarios where traditional input devices are impractical, such as the medical operating room [1]. In order to track human full-body pose in real-time, camera-based motion capture systems can be used that typically require a person to wear cumbersome markers or suits. Lately, research has focussed on markerless human pose estimation [2], [3]. However, even if multiple cameras are used, this task is challenging due to the complexity of human movements and their highly variable visual appearance in images [4], [5].

Time-of-Flight (ToF) cameras have recently created the possibility of acquiring dense, three-dimensional scans of a scene in real-time [6], without the need for expensive and complex multi-camera systems. Despite their relatively low resolution, ToF cameras are suitable for human pose estimation for several reasons. ToF cameras simultaneously generate a range image, which is almost independent of lighting conditions and visual appearance, and a grayscale intensity image, similar to conventional cameras. The provided depth information can be directly used to localize a person in front of background and to resolve pose ambiguities. Nonetheless, ToF data suffers from noise and estimating human full-body pose remains a difficult problem [7].

In this paper, we propose a method that allows tracking the full-body movements of a person from ToF images, suitable for gesture-based interaction. While learning approaches for human pose estimation ([2], [3], [8]) rely on training data and are thus restricted to a particular set of movements, our method can track general, previously unseen motions. The method is based on robustly identifying anatomical landmarks in the ToF data that then serve as targets for fitting a skeleton using inverse kinematics. We propose to represent the background-subtracted ToF depth data by means of a graph that facilitates detection of body parts. In addition, we use the optical flow between subsequent ToF intensity images for depth disambiguation when body parts occlude each other.

Representing the 3D points on the surface of a person as a graph allows us to measure geodesic distances between different points on the body. While the Euclidean distance between two body points is measured through 3D space and thus can vary significantly with body movement, the geodesic distance is defined along adjacent graph nodes, i.e. along the surface of the body. Consequently, the geodesic distance between two points on the body, e.g. the centroid and an extremity, can be assumed constant, independent of body posture [9], [10] (Figure 1). We can therefore extract anatomical landmarks by searching for points at mutual geodesic distances that correspond to the actual measurements of a person. Using only ToF depth and intensity data also decreases our depencence on visual appearance. Thus, we avoid typical problems that arise when using intensity-based feature descriptors for interest point detection, e.g. lack of texture and illumination or perspective changes.

A crucial issue is to prevent the graph constructed from the 3D points from degenerating. This can happen, for instance, when body parts occlude each other. In this case, separating the occluding body part from the part behind it becomes difficult, leading to undesired graph edges. These edges between points on different body parts result in erroneous geodesic distances, and consequently, to undetected anatomical landmarks. We address this issue by taking into account the motion occurring between subsequent frames. In particular, we identify and remove the undesired graph edges based on optical flow fields that are computed from the ToF intensity images.

Our method takes full advantage of the available information by simultaneously using ToF depth images (for segmentation and generation of 3D points), ToF intensity images (for optical flow) and the graph-based representation (for geodesic distances). Compared to other ToF-based human

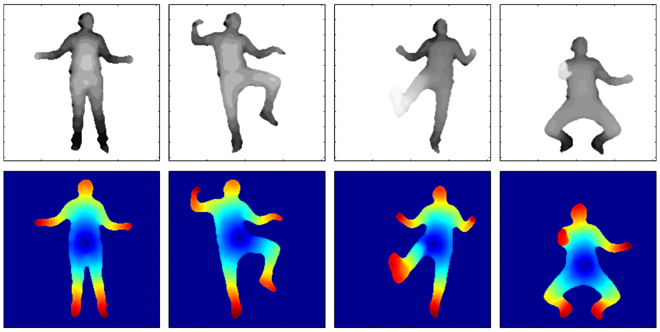

Fig. 1. Illustration of the robustness of geodesic distances against pose changes. *Top:* Background-subtracted ToF depth images for various poses. *Bottom:* Geodesic distances from the body center to all other surface points. Colors range from blue (zero distance) to red (maximal distance). Note that the distance to hands and feet remains almost constant across all poses.

body tracking approaches, ours does not require fitting body-part templates to the noisy range data. Moreover, the robust anatomical landmark detection technique allows our method to quickly recover from tracking failures. Our experiments show that we are able to efficiently track the full-body pose in several sequences containing various human movements.

II. Related Work

Techniques for human pose estimation from visual observations can be broadly categorized into learning-based approaches that facilitate the problem by means of training data (e.g. [2], [3], [11]) and approaches estimating human pose parameters from observed features without prior knowledge. A disadvantage of the former is that pose estimation is typically restricted to a set of activities known to the algorithm beforehand. For instance, Jaeggli *et al.* [3] use monocular images and extract human silhouettes as an input to a pose estimator trained on walking and running. In [11], body poses for a pre-determined activity are predicted from voxel data obtained from a 3D reconstruction system.

Methods that do not use prior knowledge for pose estimation (e.g. [4], [12], [5], [13]) are typically more dependent on reliable feature extraction, as the appearance of the human body is heavily affected by illumination and pose changes, and by noise in the observation data. Moreover, efficient state inference techniques are required to deal with the high dimensionality of full-body pose space. Kehl and van Gool [4] cope with these issues by using a multi-camera setup and generating 3D volumetric reconstructions for human pose estimation. In [13], body poses are estimated by assisting a multi-camera system with inertial sensors attached to the human body.

To overcome the limitations of visual observations, several authors have recently used ToF cameras for analysis of human motions. In [1], a system is described that recognizes simple hand gestures for navigation in medical imaging applications. The method of Jensen *et al.* [14] allows tracking the movement of legs in side-views for medical gait analysis. Holte *et al.* [15] propose a method that integrates ToF range and intensity images for human gesture recognition. Their approach is not used for pose tracking, but is able to classify upper-body gestures, such as raising an arm. The authors avoid identifying anatomical landmarks by using a global pose descriptor. Zhu *et al.* [12] present a full-body pose estimation system that relies on fitting templates for each body part to the ToF data. In [16], the authors combine a template fitting technique based on dense point correspondences with an interest point detector for increased robustness. While the approach can track full-body motion, it relies on an independent, heuristic treatment for each body part. In [8], a ToF-based method is described that simultaneously estimates full-body pose and classifies the performed activity. As opposed to our method, the system can only process movements known a priori.

Similar to our approach, Plagemann *et al.* [9] use a graph representation of the 3D data for detection of anatomical landmarks. Their technique extracts interest points with maximal geodesic distance from the body centroid and classifies them as hands, head and feet using a classifier trained on depth image patches. The method does not explicitly address the problem of self-occlusions between body parts and reportedly struggles in such situations. Without modifying the interest point detection technique, the authors add in [7] a pose estimation method embedded in a Bayesian tracking framework. Our proposed method uses optical flow measured in ToF intensity images to cope with body self-occlusions.

Optical flow has been used in [17] for motion estimation and segmentation of a person in a monocular pedestrian tracking application. Okada *et al.* [18] describe a person tracking method that combines disparity computation in a stereo setup with optical flow. Similar to our approach for disambiguation in the case of self-occlusions, an interest region map is propagated through the tracking sequence using the computed flow vectors. While this method allows tracking the bounding boxes of the head and upper body, our technique estimates the joint angles of a full skeleton body model in every frame. To the best of our knowledge, using optical flow for segmentation of occluding body parts in ToF-based human body tracking is a novel approach, enabling us to track arbitrary full-body movements.

III. Human Full-body Tracking Method

We are given a sequence $\{\mathcal{T}_t\}_{t=1}^N$ of N ToF measurements, where each $\mathcal{T}_t = (\mathbf{D}_t, \mathbf{I}_t)$ consists of a depth image $\mathbf{D}_t$ and an intensity image $\mathbf{I}_t$, both of size $n_x \times n_y$. In every frame, we initially locate L anatomical landmarks $\mathcal{P}_t = \{\mathbf{p}_i^t\}_{i=1}^L$ on a person's body, where $\mathbf{p}_i^t \in \mathbb{R}^3$, and determine the discrete landmark labels $\alpha(\mathbf{p}_i^t)$ (e.g. *head*, *left knee*). Our final goal is to estimate the full-body pose $\mathbf{q}_t \in \mathbb{R}^d$ of the person, parameterized by the d joint angles of a skeleton model.

Our method consists of an interest point detection and a model fitting part (see Figure 2). In the former, we construct a graph representation of the 3D points (section III-A) that is invariant to articulation changes and, thus, allows us to identify anatomical landmarks independent of posture (section III-B). The optical flow between the previous and the current frame, measured using the intensity images $\mathbf{I}_{t-1}$

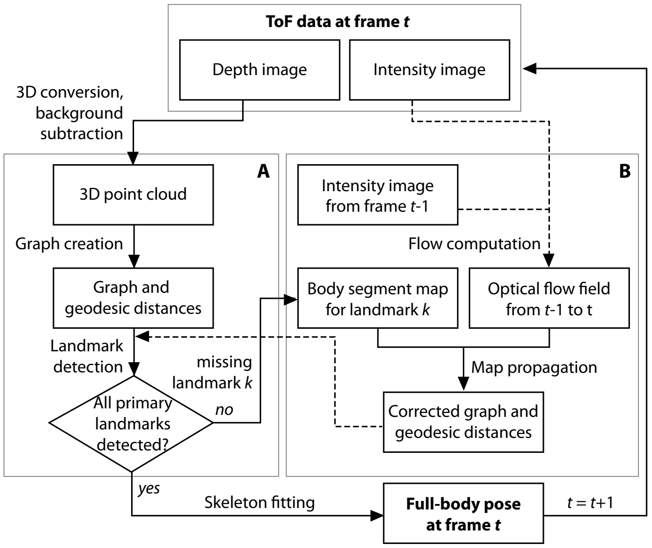
ToF data at frame t
Depth image
Intensity image
3D conversion, background subtraction
A
3D point cloud
Graph creation
Graph and geodesic distances
Landmark detection
missing landmark k
All primary landmarks detected?
no
yes
B
Intensity image from frame t-1
Flow computation
Body segment map for landmark k
Optical flow field from t-1 to t
Map propagation
Corrected graph and geodesic distances
Skeleton fitting
Full-body pose at frame t
t = t+1

Fig. 2. Schematic of the ToF-based human body tracking method. In each frame t, the algorithm constructs a geodesic distance graph based on the 3D-converted ToF depth image and extracts anatomical landmarks (A). For each undetected landmark k, the disambiguation process using optical flow on the intensity image is executed (B). The corrected geodesic distance graph for a body part k allows detecting the missing anatomical landmarks.

and $\mathbf{I}_t$, is used to track body parts that occlude each other (section III-D). We then employ a model-based skeleton fitting approach to estimate the most likely full-body pose, given the extracted anatomical landmarks (section III-E).

A. Graph-based Representation of Depth Data

Initially, we transform the depth image $\mathbf{D}_t$ into a 3D point cloud based on the known intrinsic parameters of the ToF camera, and segment the person by means of static background subtraction. Let $\mathcal{X}_t = \{\mathbf{x}_{ij}\}$ denote the resulting set of $n_x n_y$ 3D points. The notation indicates that the point $\mathbf{x}_{ij}$ corresponds to the depth image pixel with coordinates (i, j). We construct a graph $G_t = (V_t, E_t)$, where $V_t = \mathcal{X}_t$ are the vertices and $E_t \subseteq V_t \times V_t$ are the edges. Whether two vertices, i.e. 3D points, are connected with an edge or not is based on their spatial distance in 3D and their vicinity in the 2D depth image. The set of edges is defined as

$$\begin{aligned} E_t = \{(\mathbf{x}_{ij}, \mathbf{x}_{kl}) \in V_t \times V_t \mid \|\mathbf{x}_{ij} - \mathbf{x}_{kl}\|_2 < \delta \\ \wedge \|(i,j)^\top - (k,l)^\top\|_\infty \leq 1\}, \end{aligned} \quad (1)$$

where $\|\cdot\|_2$ is the Euclidean and $\|\cdot\|_\infty$ is the maximum norm and $(i, j)^\top$, $(k, l)^\top$ are the 2D coordinates of the two points $\mathbf{x}_{ij}, \mathbf{x}_{kl}$ in the depth image. For each edge $e = (\mathbf{x}, \mathbf{y}) \in E_t$, we store a weight $w(e) = \|\mathbf{x} - \mathbf{y}\|_2$. We thus connect points with a 3D Euclidean distance of less than δ that project to neighboring pixels in 2D. Incorporating the 2D neighborhood allows us to efficiently construct the graph in linear time, as opposed to computing all pairwise point distances in 3D.

Using G_t, we are able to measure geodesic distances between different body locations. The geodesic distance $d_G(\mathbf{x}, \mathbf{y})$ between two points $\mathbf{x}, \mathbf{y} \in V_t$ is given by

$$d_G(\mathbf{x}, \mathbf{y}) = \sum_{e \in SP(\mathbf{x}, \mathbf{y})} w(e), \quad (2)$$

where $SP(\mathbf{x}, \mathbf{y})$ contains all edges along the shortest path between $\mathbf{x}$ and $\mathbf{y}$. Intuitively, the geodesic distance between two locations on the body is thus the length of the shortest path over the body surface. Given a single source point, the shortest paths to all other points in the graph can be computed efficiently using Dijkstra's algorithm.

B. Detection of Anatomical Landmarks

Having constructed the graph G_t in frame t, we proceed by locating $L = 11$ anatomical landmarks $\mathcal{P}_t = \{\mathbf{p}_i^t\}_{i=1}^L$ and determining their labels $\alpha(\mathbf{p}_i^t)$. We distinguish between *primary* landmarks $\mathcal{P}'_t$ (body center, head, hands, feet) and *secondary* landmarks $\mathcal{P}''_t$ (chest, knees, elbows).

Our central assumption is that all anatomical landmarks remain at a nearly constant geodesic distance from the body center of mass, independent of body pose [9]. Detection of the *primary* anatomical landmarks therefore starts with extracting the body center of mass, given by the centroid $\mathbf{c}^t$ of the point cloud $\mathcal{X}_t$. To extract the extremities, we select all points $\mathbf{x}$ with $d_G(\mathbf{x}, \mathbf{c}^t) > \tau$. Here, τ is a person-specific threshold that approximates the distance from the body center to the shoulders (see section III-C). We therefore obtain spatially isolated sets of points that we treat as belonging to different limbs. For each of these isolated sets, we store the point with largest geodesic distance form the body center, yielding the set of primary anatomical landmarks $\mathcal{P}'_t$.

C. Initialization and Landmark Labeling

Given the locations of the primary anatomical landmarks, we need to determine their labels in order to detect the secondary landmarks, i.e. the chest, elbows and knees. For this purpose, we require a simple initialization phase where the person takes on a T-pose. Here, we measure the person-specific limb lengths and create an initial labeling of the anatomical landmarks. Each landmark $\mathbf{p}_i^0$ detected in the initialization frame ($t = 0$) is assigned an appropriate label $\alpha(\mathbf{p}_i^0)$ based on the assumed T-pose. In any subsequent frame t, we determine the labels for the primary landmarks by matching the detected positions $\mathbf{p}_i^t$ to the known landmarks in the previous frame. The label for the i-th landmark is thus

$$\alpha(\mathbf{p}_i^t) = \alpha(\bar{\mathbf{p}}^{t-1}), \text{ where } \bar{\mathbf{p}}^{t-1} = \underset{\mathbf{p} \in \mathcal{P}'_{t-1}}{\arg\min} \|\mathbf{p}_i^t - \mathbf{p}\|_2. \quad (3)$$

We can then extract the location of the *secondary* landmarks $\mathcal{P}''_t$, i.e. the chest, elbows and knees, by measuring geodesic distances from the localized primary anatomical landmarks. That is, we select points on the body as the chest, the elbows and the knees that are located at respective distances from the body center, the hands and the feet.

D. Depth Disambiguation Using Optical Flow

In cases when the extremities are clearly separated from each other, the graph-based landmark identification approach allows us to detect all primary and secondary landmarks. However, when body parts occlude each other, the graph G_t will likely contain edges that connect points on different body parts. In such a situation, two points $\mathbf{x}, \mathbf{y} \in V_t$ on distinct body parts can easily satisfy the two conditions of

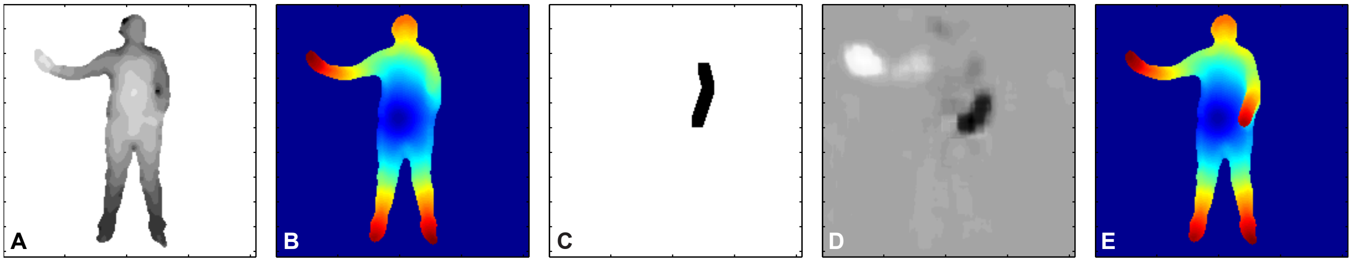

Fig. 3. Illustration of the depth disambiguation approach using optical flow. *A:* Background-subtracted ToF depth image with a hand in front of the torso. *B:* Geodesic distance map computed using the graph G_t with the origin at the body center. The occluding arm is too close to the torso for being separated. *C:* Body segment map for the arm obtained in the previous frame. *D:* Optical flow field (x-component) from previous to current frame. *E:* Geodesic distance map after removal of undesired edges in G_t. The arm is now separated and has the expected geodesic distance from the body center.

(1). Consequently, the geodesic distances will be computed inappropriately. Figure 3.B gives an example where an arm in front of the torso is connected to the upper body and the geodesic distance from the body center to the hand is underestimated. Without correction, anatomical landmarks on the arm cannot be detected.

We therefore propose a disambiguation approach that makes use of movement occurring between frames. Assuming that distinct body parts move separately, this approach allows us to disconnect points belonging to different body parts. We introduce a binary map indicating the location of the entire *occluding* body segment in the depth image. This map is propagated and updated from frame to frame using optical flow, until the body parts become separable again.

1) ***Creation of body segment map:*** Let $\mathbf{p}_m^t \in \mathcal{P}_t'$ be the location of a primary anatomical landmark at time t and let b_m^t denote the corresponding body part, i.e. an arm or a leg. We define the body segment map $\mathbf{M}_t^m$ for b_m^t to be a binary image of the same size as the depth image $\mathbf{D}_t$, such that

$$\mathbf{M}_t^m(i,j) = \begin{cases} 1 & \text{if } d_G(\mathbf{p}_m^t, \mathbf{x}_{ij}) < \mu, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

All pixel locations (i,j) in the map are assigned a value of 1 if the geodesic distance between their corresponding 3D point $\mathbf{x}_{ij}$ and the landmark $\mathbf{p}_m^t$ does not exceed μ. This threshold is chosen based on the length of the person's limbs (determined during initialization), such that the entire body segment b_m^t is included in the segment map. Figure 3.C shows a body segment map for the person's left arm.

2) ***Map propagation using optical flow:*** For every primary landmark $\mathbf{p}_m^t$ that is not detected using the approach described in section III-B, we obtain the corresponding body segment map $\mathbf{M}_{t-1}^m$ from the previous frame. If the landmark was detectable in that frame, we construct the map according to (4), otherwise we assume that the map is available from previous propagation steps. Let $\mathcal{F}_t = (\mathbf{F}_{t,x}, \mathbf{F}_{t,y})$ denote the optical flow between the ToF intensity images $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$. $\mathbf{F}_{t,x}(i,j)$ is the x-component of the estimated movement for pixel (i,j) between the two images, and similarly for y. Figure 3.D shows an exemplary flow field. We use the optical flow to update the map $\mathbf{M}_{t-1}^m$ to reflect the assumed position of body part b_m^t in frame t. The propagated map $\mathbf{M}_t^m$ is computed such that

$$\mathbf{M}_t^m(i + \mathbf{F}_{t,x}(i,j), j + \mathbf{F}_{t,y}(i,j)) = \mathbf{M}_{t-1}^m(i,j). \tag{5}$$

A set of image processing steps are applied to the propagated map, including morphological operations, to remove noise and cavities caused by artefacts in the optical flow field.

3) ***Removal of undesired graph edges:*** Using the updated and corrected map, we can remove the undesired edges in the graph G_t that connect points on body segment b_m^t to the body part in the background, e.g. the torso. We update the set of edges as $E_t = E_t - F$, with

$$F = \{(\mathbf{x}_{ij}, \mathbf{x}_{kl}) \in E_t \mid \mathbf{M}_t^m(i,j) \neq \mathbf{M}_t^m(k,l)\}, \tag{6}$$

where $\mathbf{x}_{ij}$ is the 3D point corresponding to the location (i,j) in the body segment map. In other words, all edges are removed where one point lies within the body segment map and the other point does not. Figure 3.E illustrates the geodesic distances from the body center after the edges between the occluding arm and the torso have been disconnected. The corrected graph allows us to identify the primary and secondary anatomical landmarks on body segment b_m^t by re-computing the geodesic distances from the body center and selecting points with a maximal distance, as described in section III-B.

In the situation that multiple primary anatomical landmarks cannot be detected, we repeat the process described above for every missing landmark, each time propagating the appropriate body segment map and disconnecting undesired graph edges. Note that the map propagation step, although based upon optical flow between subsequent frames, does not fail without movement. In such cases, the optical flow field is close to zero and the body segment map simply remains unchanged.

E. Skeleton Fitting Using Inverse Kinematics

Once the anatomical landmarks $\mathcal{P}_t$ have been identified and labelled in frame t, we estimate the full-body pose parameters $\mathbf{q}_t \in \mathbb{R}^d$ by fitting a skeleton to the detected points. Our skeleton model consists of 16 joints, distributed over five kinematic chains (both arms and legs, torso), where individual joints have one, two or three degrees of freedom. In total, the parameter space for the skeleton model has $d = 38$ dimensions.

Starting with the torso chain that is registered in the body centroid $\mathbf{c}^t$, the full-body pose is determined, intuitively, by attracting selected joints of the kinematic chains (effectors) to the locations of the anatomical landmarks (targets). Formally, the objective is to find the optimal joint angle configuration $\mathbf{q}_t$ such that the residual error

$$\mathcal{E}(\mathbf{q}, t) = \sum_{i=1}^{L} \|\mathbf{p}_i^t - f_i(\mathbf{q})\|_2^2 + c(\mathbf{q}), \tag{7}$$

is minimized. Here, $f_i : \mathbb{R}^d \rightarrow \mathbb{R}^3$ is a forward kinematic function that computes the 3D position of the i-th joint, given a vector $\mathbf{q}$ of joint angles. We add a term $c(\mathbf{q})$ that penalizes joint angle configurations violating a set of constraints. The term $c(\mathbf{q})$ increases polynomially when any of the joint angles approaches its pre-specified lower and upper limits.

To find $\mathbf{q}_t$, we employ an iterative Gauss-Newton optimization approach that, starting with an initial value $\hat{\mathbf{q}}_0$, computes updates $\Delta_{\mathbf{q}}$ such that $\hat{\mathbf{q}}_{i+1} = \hat{\mathbf{q}}_i + \Delta_{\mathbf{q}}$, until convergence. In each frame, we use the joint angles of the previous frame as an initial value, $\hat{\mathbf{q}}_0 = \mathbf{q}_{t-1}$. Assuming incremental body movement between subsequent frames, this increases convergence rates and decreases the probability of hitting local minima.

IV. Experiments and Results

In order to evaluate our ToF-based body tracking method, we recorded a series of 20 testing sequences using a PMD-Vision CamCube ToF-camera with a resolution of 204×204 pixels. Each of the sequences consists of around 400 frames, at a frequency of 10 Hz. The recorded movements range from simple motions, such as waving an arm, trough complex full-body movements with occlusions between body parts. Figure 5 gives an overview of the movements in our training set. To provide a quantitative assessment of our method, we simultaneously recorded ground truth data with a marker-based motion capture system that was synchronized to the ToF camera and registered to its coordinate frame. Motion capture markers were placed on the back of the person to prevent interference with the ToF measurements.

In our experiments, the ToF depth and intensity images were pre-processed using a median filter to decrease the level of noise. We segmented the person from the background in each frame by subtracting a static depth image of the lab acquired beforehand. After constructing the graph G_t in every frame, geodesic distances to all body surface points were precomputed and stored in a distance map, similar to the maps in Figure 1. We used the Horn-Schunck method for computation of the optical flow fields and low-pass filtered each of the spatial flow field components. Our current Matlab implementation reaches tracking rates of 2-4 frames per second.

A. Precision of Full-body Pose Estimation

The motion capture system provides the 3D positions of the $K = 16$ body joints $\{\mathbf{s}_i\}_{i=1}^K$ of the skeleton model described in section III-E. As an error metric, we therefore compute in every frame the average Euclidean distance between the estimated and true locations of these body joints. We define the distance error as

$$e_{\text{dist}}(t) = \frac{1}{K} \sum_{i=1}^{K} \|\mathbf{s}_i - \hat{\mathbf{s}}_i\|_2, \tag{8}$$

where $\hat{\mathbf{s}}_i$ is the estimated 3D position of the i-th skeleton joint and $\mathbf{s}_i$ is the corresponding ground truth. Note that, even in the case of a perfectly estimated full-body pose, $e_{\text{dist}}(t)$ will not be zero, since the markers of the motion capture system do not coincide with our detected anatomical landmarks. Moreover, the motion capture system fits the skeleton to assumed locations of joints within the body, whereas our fitting targets are on the body surface.

Averaged over all testing sequences, our method achieved a distance error of $\bar{e}_{\text{dist}} = 70.1$ mm with a standard deviation of 9.8 mm. Figure 4 shows plots of the distance error over the length of two typical testing sequences. The overlaid full-body pose prediction and ground truth for selected frames allows for a better interpretability of the results. The left graph in Figure 4 corresponds to one of the easy sequences where only the arms are moved, however, including body self-occlusions. In this case, the distance error averaged over all joints is around 50 mm. The maximum error in each frame rarely exceeds 100 mm. On the right side, results are shown for a more difficult sequence including full-body movement. Especially when legs are raised, the average error increases to around 100 mm. The effect of the maximal value of 250 mm for individual joints is visualized in example 7 (Figure 4), where the position of the right knee deviates from the ground truth. This being an example for worst-case deviations, our method compares favorably to current state-of-art methods for ToF-based full-body pose tracking (e.g. [7]).

B. Qualitative Assessment

Figure 5 provides example images from our testing sequences for a qualitative assessment of our method. As can be seen, the estimated full-body poses match with the ToF depth images. Poses are predicted faithfully, even in cases where arms or legs move towards or away from the camera. In particular, the situation where a hand is stretched forward and occludes the arm itself is handled well. The second row of images in Figure 5 shows cases where one or both hands move in front of the torso. Here, our method relies on the optical flow-based disambiguation approach described in section III-D. Tracking does not fail, even when more than one limb moves in front of the body. The speed of movements is not a critical parameter to our technique, as long as the positions of primary anatomical landmarks can be matched sucessfully accross subsequent frames. In our experiments, tracking problems mainly occurred when two arms or legs crossed each other in front of the torso. In such cases, parts of one limb were cut off by the occluding limb, resulting in inaccurate landmark detections.

V. Discussion and Conclusion

We have presented a method for tracking human full-body pose from sequences of ToF camera images. The approach

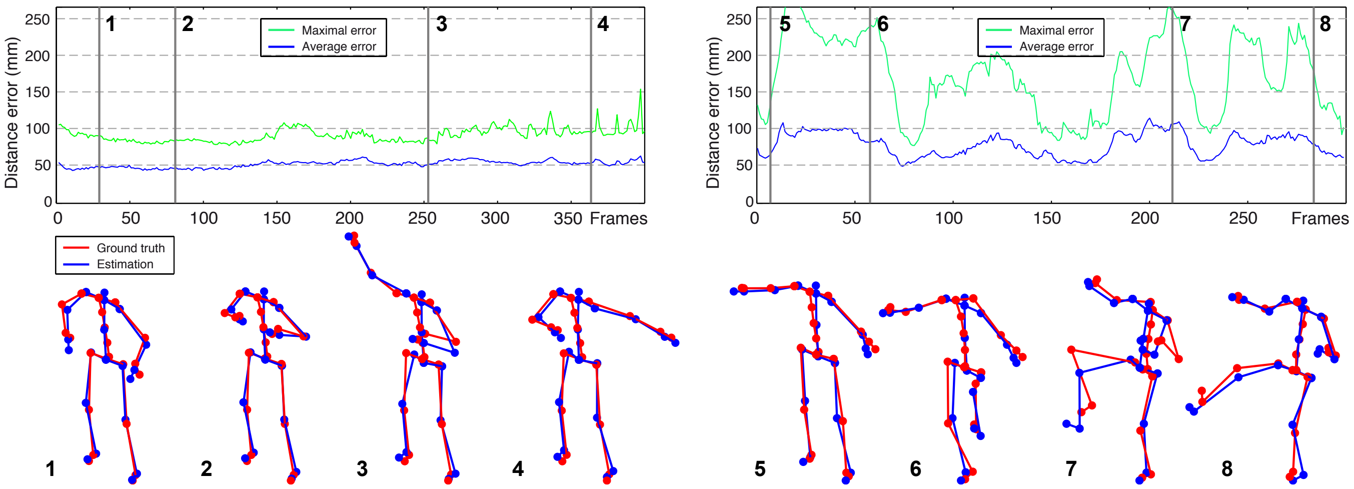

Fig. 4. Illustration of quantitative pose estimation results. The two graphs show the distance error $e_{\mathrm{dist}}(t)$ over the length of two exemplary testing sequences. The average error over all joints is plotted (blue), along with the maximum error in each frame (green). *Left:* Typical sequence where only hands are moved, including self-occlusions. *Right:* Typical sequence involving full-body movement. Results for selected frames are visualized below with overlaid estimated (blue) and ground truth poses (red).

does not require any training data and is able to track arbitrary movements. Initialization is limited to holding a T-pose, while the approximate limb lengths of the person are measured. This step can be overcome by adapting automatic body calibration methods, such as [19], to the present setting. While the current implementation is close to providing real-time frame rates, we believe there is sufficient potential for improving computational efficiency, without requiring GPU acceleration. An inherent assumption of our method is that persons are facing the ToF camera and do not fully rotate around their vertical axis. We argue that this assumption is reasonable for gesture-based human-machine interaction.

Our method takes full advantage of the data provided by ToF cameras by utilizing both, depth and intensity information. Based on the depth data, we segment the person in front of static background and construct a graph-based represenation of the 3D points. This graph allows us to robustly identify anatomical landmarks in each frame by selecting points with a maximal geodesic distance from the body center of mass. In cases where body parts occlude each other, we rely on optical flow, computed on the ToF intensity images, to disconnect occluding limbs from the body part behind. The experimental evaluation presented in this paper shows that our method can track various full-body movements, including self-occlusions, and estimate 3D full-body poses with a high accuracy.

REFERENCES

[1] S. Soutschek, J. Penne, J. Hornegger, and J. Kornhuber, “3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras,” *Computer Vision and Pattern Recognition Workshops*, Apr 2008.

[2] R. Urtasun and T. Darrell, “Sparse probabilistic regression for activity-independent human pose inference,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jan 2008.

[3] T. Jaeggli, E. Koller-Meier, and L. V. Gool, “Learning generative models for multi-activity body pose estimation,” *International Journal of Computer Vision*, vol. 83, no. 2, pp. 121–134, 2009.

[4] R. Kehl and L. Gool, “Markerless tracking of complex human motions from multiple views,” *Computer Vision and Image Understanding*, Jan 2006.

[5] J. Bandouch, F. Engstler, and M. Beetz, “Accurate human motion capture using an ergonomics-based anthropometric human model,” *Articulated Motion and Deformable Objects (AMDO)*, Jan 2008.

[6] A. Kolb, E. Barth, R. Koch, and R. Larsen, “Time-of-flight sensors in computer graphics,” *EUROGRAPHICS*, pp. 119–134, 2009, notizen.

[7] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, “Real time motion capture using a single time-of-flight camera,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[8] L. A. Schwarz, D. Mateus, V. Castaneda, and N. Navab, “Manifold learning for tof-based human body tracking and activity recognition,” *British Machine Vision Conference (BMVC)*, pp. 1–11, Aug 2010.

[9] C. Plagemann, V. Ganapathi, and D. Koller, “Real-time identification and localization of body parts from depth images,” *IEEE International Conference on Robotics and Automation (ICRA)*, Jan 2010.

[10] M. Mortara, G. Patane, and M. Spagnuolo, “From geometric to semantic human body models,” *Computers and Graphics*, vol. 30, pp. 185–196, Mar 2006.

[11] Y. Sun, M. Bray, A. Thayananthan, B. Yuan, and P. Torr, “Regression-based human motion capture from voxel data,” *British Machine Vision Conference (BMVC)*, 2006.

[12] Y. Zhu, B. Dariush, and K. Fujimura, “Controlled human pose estimation from depth image streams,” *Computer Vision and Pattern Recognition Workshops*, 2008.

[13] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn, “Multisensor-fusion for 3d full-body human motion capture,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, May 2010.

[14] R. Jensen, R. Paulsen, and R. Larsen, “Analyzing gait using a time-of-flight camera,” *Scandinavian Conference on Image Analysis*, pp. 21–30, 2009.

[15] M. B. Holte, T. B. Moeslund, and P. Fihl, “Fusion of range and intensity information for view invariant gesture recognition,” *Computer Vision and Pattern Recognition Workshops*, May 2008.

[16] Y. Zhu and K. Fujimura, “A bayesian framework for human body pose tracking from depth image sequences,” *Sensors*, May 2010.

[17] S. Denman, V. Chandran, and S. Sridharan, “An adaptive optical flow technique for person tracking systems,” *Pattern recognition letters*, vol. 28, no. 10, pp. 1232–1239, 2007.

[18] R. Okada, Y. Shirai, and J. Miura, “Tracking a person with 3-d motion by integrating optical flow and depth,” *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, Sep 2000.

[19] J. F. Obrien, B. Bodenheimer, G. Brostow, and J. Hodgins, “Automatic joint parameter estimation from magnetic motion capture data,” *GRAPHICS INTERFACE*, Jan 2000.

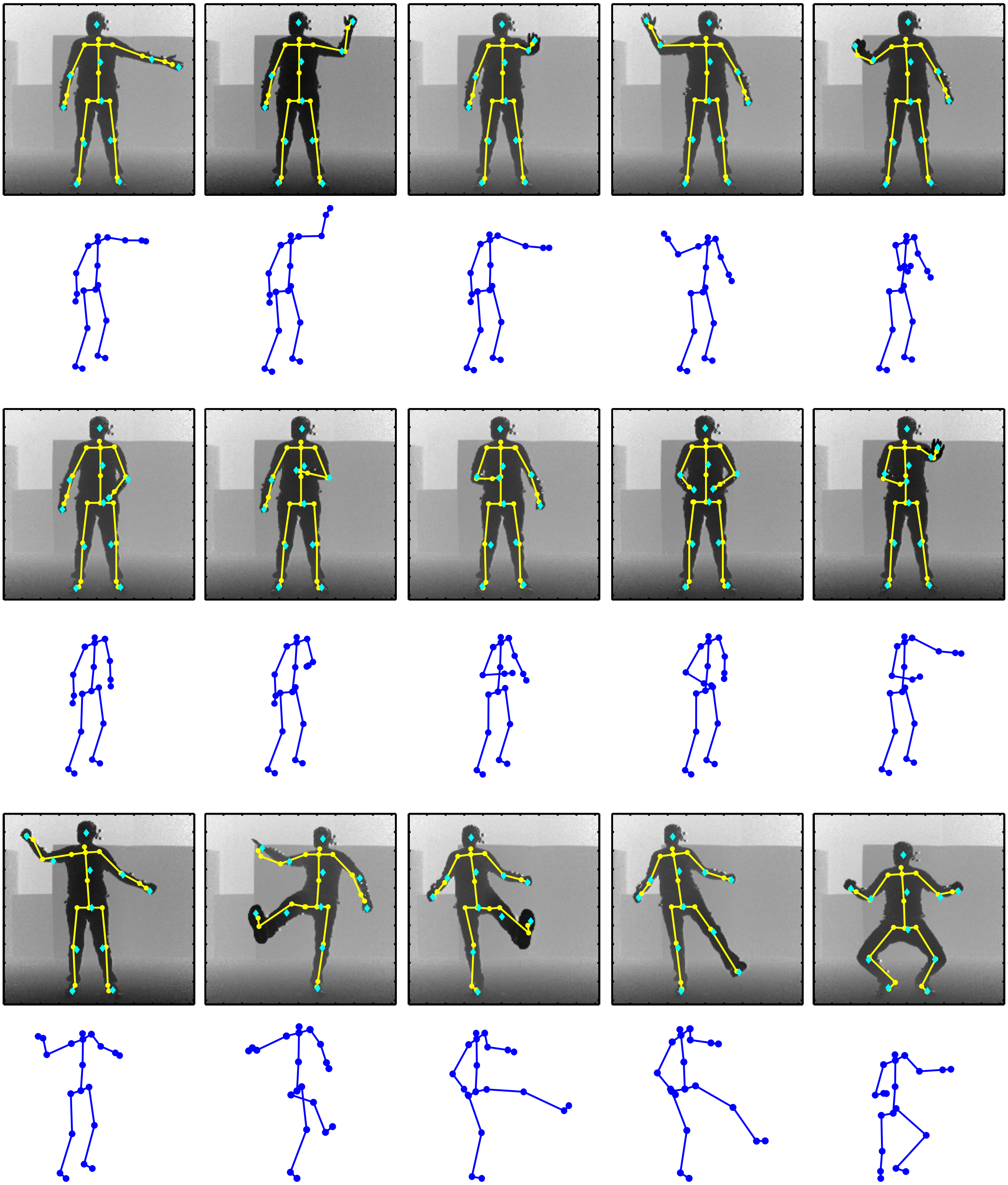

Fig. 5. Qualitative assessment of the proposed full-body pose estimation method. In each of the three rows, ToF depth images are shown, overlaid with projections of the estimated skeleton pose (yellow). Blue markers indicate the positions of detected anatomical landmarks that play the role of targets for skeleton fitting. Below each row, perspective views of the corresponding estimated poses are displayed, emphasizing the 3D appearance of the predictions.