Towards view-invariant Gait Recognition from monocular video based on Human Pose Estimation

Master thesis report, April 2021

Bousias Dimitrios

***Abstract*—Gait recognition is the biometric method that can differentiate and identify individuals by the way they walk. Gait as a biometric feature has some interesting characteristics as it can be collected at distance while it can be very hard to fake. Previous gait recognition works that rely on human silhouette representation, are often dependent on robust contouring or background extraction methods. Additionally they can be very limited regarding the viewing angle or require specific conditions to be met. Inspired by recent progress in the field of human pose estimation and skeleton based gait recognition methods, we propose a framework for extracting markerless motion capture data from monocular video and identifying individuals based on extracted features. The generalizing power of off-the-shelf pose estimators towards in the wild videos is tested. The approach is aiming towards a view-angle and clothing invariant solution. A gait dataset is acquired to validate the uniqueness and permanence of various gait features. We report results of verification and identification experiments which are compared to respective ones attained with a commercial depth sensor. Correct identification rates of 79% up to 88% in the overall experiment are achieved using different combinations of features and template matching methods. Possible shortcomings of the method related to view-angle dependent bias or the filtering of identity information are discussed.**

I. Introduction

Gait recognition can be associated with the task of recognizing or identifying individuals based on characteristics or patterns exhibited in ones way of walking, typically from a distance. It seems to be generally accepted that each person has a unique style of walking which can be attributed to their physical characteristics and properties among other factors. Since it can be challenging to record all kinesiological parameters that can form a basis for identification, gait recognition approaches have to rely on video sequences taken in controlled or uncontrolled environments. Even if we are able to measure certain gait parameters with high accuracy, we are still not certain if the knowledge of these parameters can provide adequate discriminating power to enable gait recognition technologies at large scale [1]. Gait recognition systems can face further limitations regarding the reliable extraction of discriminative features since gait itself can change over time and can be affected by factors such as clothes, footwear, walking surface, walking speed, emotional state [2] and medical conditions.

We approach the problem in a model-based manner, where the computer vision and biometric recognition aspects are disentangled. A model of the human skeleton and recordings of the movements of its various parts, can offer a suitable representation on which to base our recognition experiments. To that end we utilize human pose estimation methods to encode rich visual features into pose and movement patterns in a skeleton representation, which can allow us to perform biometric experiments in data of much lower dimensionality.

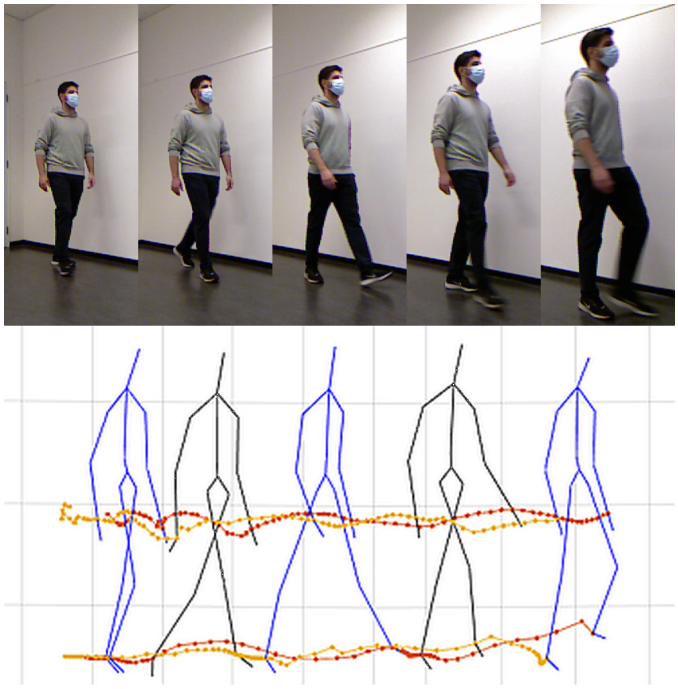

Fig. 1. (Above) Frames from a human gait sequence, (Below) Skeleton representation of gait using a marker-less, monocular motion capture approach powered by pose estimators.

Human pose estimation(HPE) is the process of estimating the configuration of the body (pose) in human depicting images. It can be performed both on single images and on video or sequences of images. The goal of the task can be to determine the position of human body joints as a two-dimensional(2D) [x,y] or three-dimensional(3D) [x,y,z] vector of Cartesian coordinates. In the former case joint positions are typically determined in terms of pixel position (width,height) within an image frame, while in the latter case a 3D vector of position values is determined in a camera coordinate or world coordinate system.

By estimating and keeping track of pose over time one can

achieve a detailed description of human movement. To that extend HPE approaches provide a visual based solution to Motion Capture without the need of elaborate studio setups and body suits with identifying markers. To the same end, methods of capturing gait signals with wearable inertial sensors were deemed rather inaccesible/impractical and not suitable to our needs, so they were not considered.

The availability of large annotated datasets and the introduction of deep reinforcement learning methods proved vital for progress in the field of HPE. Latest works (monocular [3], multi-view [4]) report estimation errors in the range of millimeters to few centimeters from ground truth. This appears very promising and opens room for exploration for the various applications of HPE including gait recognition.

Motivated by this we wish to research the development of a free-view gait recognition approach that utilizes HPE to be used on monocular videos in the wild. Though this has to be based on some strong **assumptions**:

- The methods used are able to properly generalize in order to handle never-seen-before videos.
- The methods are able to reliably extract skeleton representations without strong biases regardless of viewing angle.
- The resulting representations maintain their discriminative capabilities.

The model based approach that we employ utilizes detections of human parts. Their representation as points in space which is expressed in terms of coordinates is referred as keypoints. In our method we make use of 2D keypoints and map them into 3D space. For this reason we refer to the keypoints extracted this way as Keypoints2D3D. Keypoints extracted from depth data using the Kinect sensor(more on methodology section) are referred as KeypointsKinect. We proceed to formulate the basic **Research Questions** that this work will investigate.

Main Research Question: *Can a 3D gait recognition model-based method for monocular RGB, using 2D joint keypoints from arbitrary viewing angle, be realized?*

Subquestions:

1) What 2D keypoint extraction methods exist in literature?
   - Which one is best fitting to this problem?
2) What methods do exist for conversion of 2D keypoints to 3D keypoints?
   - Which works best in our case?
3) What kind of data is available? Does it fit our needs?
4) Quality of extraction: What percentage of 2D joint keypoints can be detected throughout a sequence (per viewing angle)?
5) How does accuracy/reliability of Keypoints2D3D vary with viewing angle?
   - How to properly measure the accuracy of extracted keypoints ?
   - What is the accuracy obtained? -for different subjects, -for different viewing angles.
6) What are useful features to be used from 3D Keypoints to allow identification of persons?
   - What kind of features can be extracted from KeypointsKinect (from commercial RGB-D sensor used for comparison) for Biometric Identification?
   - Can the same be used for Keypoints2D3D? How do they compare? Are they robust to view angle change?
7) What is the performance of a system using those features on suitable data (our own collection of gait sequences)?

In section 2 we focus on related studies and bibliography. Recent works but also works that paved the way to current state-of-the-art approaches are presented both for the HPE and Gait recognition part. A short description of relevant large datasets and benchmarks required to train the deep-learning approaches follows. Section 3 contains the methodology of our approach and justification for our decisions. All aspects of our pipeline framework are explained in detail. In section 4 the experimental setups needed to evaluate our approach and its various aspects are explained. The acquisition process of gait sequences from multiple subjects (to facilitate those experiments) is described. The next sections cover qualitative and quantitative results of the process and provide commentary. Successes, fails , reasoning and possibilities for improvement are discussed. The final section offers general conclusions and lays-out pathways for necessary follow-up research works in the field.

II. Related Work

The use of image based observations, to recover the pose of an articulated body which consists of joints and rigid parts is a method that has been widely used in algorithms and systems for articulated body pose estimation.

A classical approach to the problem is making use of the pictorial structures framework [5] [6]. This idea aims to represent an object (and a human body as such) as a collection of ”parts” that can be arranged in a non-rigid (deformable) configuration. Each part corresponds to an appearance template that has to be matched in an image. The spatial connection between parts can be seen as a spring that exhibits certain degrees of freedom but also follows certain restrictions. When all parts have been parametrized in terms of location(pixels) and orientation the resulting structure is a model of the body's articulation and pose by extension. The idea can also be expanded to 3-dimensional structures to apply on multi-view systems [7]. The limitation of the method comes from having models that do not necessarily depend on image data and therefore having to constantly enrich the representational power of such models. In their work [8] Yang and Ramadan expand on the idea by expressing complex joint relationships with the use of a mixture model of parts. They create collections of templates arranged in a deformable configuration that they call deformable part models. Each model contains both global and part templates which are then matched in an image in order to recognize an object or body.

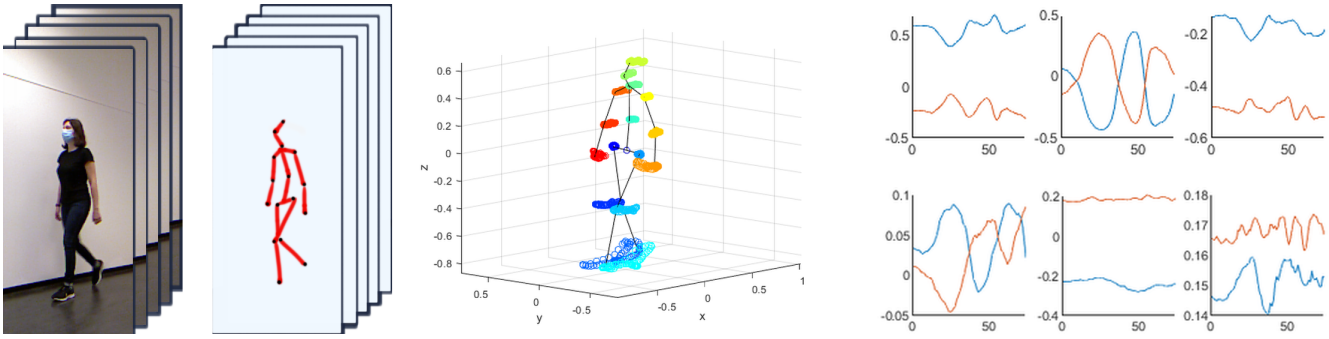

Fig. 2. Data flow in the process. A sequence of RGB frames is converted into frames of 2D keypoints, which in turn are lifted to 3D. Positional and angular signals are extracted as gait specific features to be used for recognition purposes.

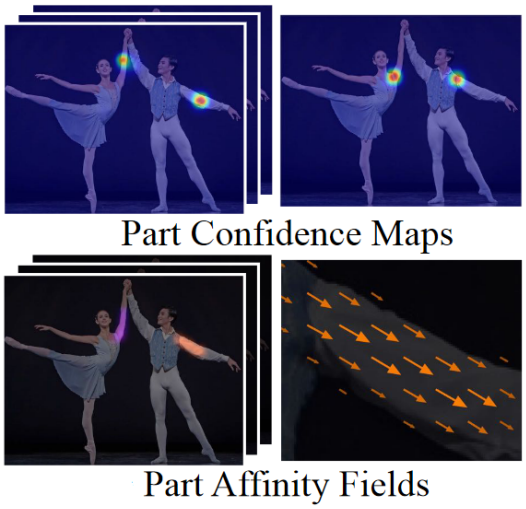

Fig. 3. Predictions from the work in [11]. On top heatmaps of confidence for part localization. On bottom vector field noting the association of a pair of parts

In recent years Convolutional Neural Networks have been almost universally adopted as the main building block of pose estimation systems. These works can be grouped into 2D and 3D systems. Below some of the most influential works are briefly described.

Deep Learning methods for 2D pose estimation

In DeepPose [9] the pose estimation is formulated as a regression problem of the body joints. At its core the model used a modified AlexNet [10] backend with an added final layer that outputs x,y joint coordinates. An L2 loss for regression is used to train the model. The authors argue about the problem in a holistic fashion (Given a holistically estimated pose even certain occluded or hidden joints can be estimated). A cascade of CNN-based regressors is used to refine an initially coarse pose and improve estimates.

The approach in [12] introduces the idea of Heatmaps, which are mappings of the confidence of the model for the location of joint-keypoints at each pixel. A multi-resolution CNN architecture is used to run an input image through multiple resolution banks in parallel and thus capture features at various scales. The joint use of a graphical model alongside the main ConvNet is proposed to learn typical spatial relationships between the joints. A Mean Squared Error(MSE) distance is used for training. It encodes the difference of the predicted heatmap to the target heatmap (The target can be a 2D Gaussian of constant variance that is centered at the ground truth keypoint location). Another interesting heatmap based model was proposed in [13]. It introduced a novel convolutional network architecture of repeated bottom-up and top-down processing with intermediate supervision.It is referred to as a Stacked-hourglass network because of the successive processes of pooling and upsampling that are performed to produce the final predictions.At each stage a convolution layer and max-pooling layers are used for feature processing, and residual connections are used to propagate results of previous stages.

Convolutional Pose machines introduced in [14] try to learn spatial relationships with the use of gradually increasing receptive fields. It is a multistage architecture consisting of an image feature computation module that is followed by a prediction module, that predict heatmaps(belief maps in the paper). Since it is completely differentiable it can be trained end to end, although the authors opt for intermediate supervision as well to handle vanishing gradients. The architecture can increase in complexity by stacking repetitions of the prediction module.The number of stages utilized can be treated as a hyperparameter on its own.

Top-down approaches can suffer from inaccurate bounding boxes or clutter, especially in multi-person or crowded images. A human pose has to be fit within the specified boundaries which can cause inaccuracies. Fang et al. in [15] (open sourched as Alpha Pose [16]) try to tackle inaccurate bounding boxes of human detectors and pose proposals,using an elaborate scheme.

OpenPose [11] [17] takes a bottom up approach with a multi-stage CNN architecture.Non parametric representations called Part Affinity Fields are used to encode part-to-part assosiation.They are essentially learnable vector fields that map spatial features to body part connections.This work is the original winner of the COCO keypoint challenge 2016 [18]. It can perform Multi-person estimation in real-time.

Two-dimensional pose estimation can be considered as previously mentioned a specialization of the more generic Object detection task. Mask R-CNN [19] showcased state of the art results while also providing pixel precise object masks (instance segmentation) in a single model. Those masks can be used to provide utility for various other tasks in pipeline configurations. In [20] Xiao et al. proposed a model that outperformed most of the previous works while aiming for a simpler architecture. It consists of a ResNet [21] backbone with the addition of some deconvolutional layers at the end. Their system works in a top-down fashion and utilizes optical flow and a greedy strategy to track poses across video frames. While most of the previous papers take a high-to-low-to-high representation approach, HRNet from Sun et al. [22] maintains a high res representation throughout the process. The model consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution modules(multi-scale fusion). This work is currently the top performing in the tasks of Keypoint Detection and Single/Multi-person pose estimation in the COCO dataset.

3D pose estimation

Recent methods for 3D human pose estimation from RGB images can be grouped into two main categories based on their training pipelines. In the first category a convolutional neural network is typically trained to estimate the 3-dimensional pose directly from the input images. Pavlakos et al. in [23] integrate the volumetric representation using a coarse-to-fine supervision method to directly predict 3D volumetric heatmaps. Dabral et al. in [24] proceed to create a weakly-supervised ConvNet estimator of human pose and propose illegal-angle loss and a symmetry loss for the training of their network. In [25] Sun et al. propose an effective integral regression approach that tries to unify the heatmap representation and joint regression approaches. Some recent works are focusing on fitting parametric meshes or other morphable models on top of the human body. Kanazawa et al in [26], Guler et al. in [27] present end to end CNN frameworks that reconstruct 3-dimensional meshes of the human body from a single RGB frame. Models that work directly on RGB images can capture rich context information contained in images of human poses. That said, the lack of intermediate features and supervision causes the final 3D pose to be affected by factors such as the image's background, lighting and the depicted person's clothing among others.

In the second category of approaches the authors construct 3D joint estimation models that are built on top of high-performance 2D keypoint detectors. The process of inferring the 3D joint locations from 2D keypoints is known as "lifting". Tome et al in [29] propose a strategy of iteratively using 3D to 2D projections and vice versa to improve predictions in a stage-like manner. They make a case of giving strong emphasis on preprocessing training data poses to eliminate ground plane rotation, left-right symmetry etc. In [30], one of the less complex approaches, Zhao et al. make use of a dense fully connected network to efficiently lift the 2D keypoints into the 3D space while also tackling noise and missing data problems. The work in [31] by Chen el al. regards the 3dimensional pose estimation as a matching problem and try to find the best matching 3D pose of the 2D keypoints input by a nearest-neighbor model. Even though it is a simple solution it outperformed most other methods at the time. In [32] Martinez et al. proposed a fully-connected residual network to effectively regress the 3D joint locations from 2D keypoint inputs. Lee et al in [33] introduced a framework that makes use of LSTMs to reconstruct the depth from the centroid of the pose to edge joints. In a different work Chen et al. in [34] presented a weakly-supervised method for learning a representation that is geometry-aware to try and bridge multi-view images for the task of pose estimation.The approaches that make use of such an "image-2D-3D" pipeline seem to outperform the end-to-end solutions mentioned above. That can be attributed to the fact that the 2D detectors can be trained on large scale datasets of indoor and outdoor images with 2D annotations that are readily available. Making use of this strong intermediate feature appears to make the 3D estimation models more robust.

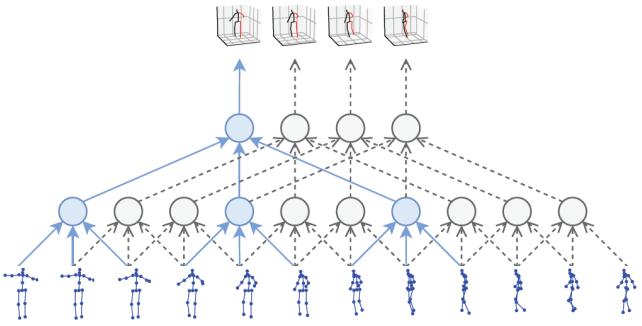

Fig. 4. Schematic showcasing temporal convolutions approach in [28]. A sequence of skeletons in 2D contribute to produce a temporally smooth sequence of 3D skeletons. The use of dilated convolutions allows for a bigger receptive field.

Video-based Approaches: Another type of approach such as [35] by Hossain et al., make use of video or group of frames inputs to leverage temporal consistency and produce smooth joint trajectory sequences. Pavllo et al. [28] propose temporal convolutions on 2D keypoints with dilated stride to reach similar results. Chen et al. [36] build on that idea to propose an Anatomy-Aware framework. Instead of directly regressing the 3D joint locations, they decompose the task into bone direction prediction and bone length prediction, from which the 3D joint locations can be completely derived. Their main motivation is the fact that the bone lengths of a human skeleton (should) remain constant across time. In one of the latest works Cheng et al. [37] build a pipeline solution that relies on 2D confidence heatmaps and occlusion annotations. Since the latter are not available on the common datasets they propose a "Cylindrical-man" model to approximate the occupation of body parts in 3D space and derive perceived occlusions.

Datasets for Human Pose estimation

The *MPII human pose dataset* is a multi-person 2D Pose Estimation dataset comprising of nearly 500 different human

activities, collected from Youtube videos. MPII was the first dataset to contain such a diverse range of poses and the first dataset to launch a 2D Pose estimation challenge in 2014.

The *COCO keypoints dataset* is a multi-person 2D Pose Estimation dataset with images collected from Flickr. COCO is the largest 2D Pose Estimation dataset, to date, and is considered an important benchmark for testing 2D Pose Estimation algorithms.

Human3.6M is a single-person 2D/3D Pose Estimation dataset, containing video sequences in which 11 actors are performing 15 different activities were recorded using RGB and time-of-flight (depth) cameras. 3D poses are obtained using 10 markered MoCap cameras. Human3.6M is the biggest 3D Pose Estimation dataset with real(non-synthetic) images, to date.

Gait Recognition

Video gait recognition methods can be mainly categorized based on the representation of gait and the data they work on to extract valuable features

Silhouette based methods: In the first category of works [43] [44] [45] authors choose a silhouette representation of gait images. Silhouettes are typically extracted with the use of edge detectors and background removal procedures or with an image segmentating network [46], and a binarization of the result. A Gait Energy Image (GEI) produced as an average of all silhouettes in a gait sequence, or Gait Entropy Image (GEnI) created by calculating the entropy of pixels in the sequence are used as features in the recognition process. Methods working with the described framework have shown promising results but are heavily limited by covariate factors of gait such as overall appearance and clothing and are very dependent on a robust silhouette extraction. Additionally those methods are usually limited into recognizing sequences captured from the same camera position or viewing angle.

Gait Recognition from Skeleton Data: An other category of works focuses on skeleton representations of captured gait sequences. Those skeletons can be derived from MotionCapture Data [47], depth sensors (typically commerically available Kinect sensors [48], or pictorial structures [49] [50]. Authors in [51] use skeletons containing joint positions extracted with a 2D pose estimator, althought their results are not totally robust to angle view changes. In PoseGait [52] a workflow similar to ours is followed where spatio-temporal features of skeleton sequenses are calculated based on a 3D skeleton representation. A CNN architecture follows that is able to extract higher level features used for recognition purposes. This approach can be robust to view changes but it requires a multi-camera setup which can not always be practical.

In some of the Skeleton based methods, where we draw inspiration from, handcrafted features are directly extracted from the joint positions. Those features are typically human-interpretable which can help with the understanding of where the discrimative power of gait comes from. Those features are usually characterized as Static or Dynamic depending on whether they change throughout a sequence. The most prominent works and some details of the approaches are presented on the table above.

III. Methodology

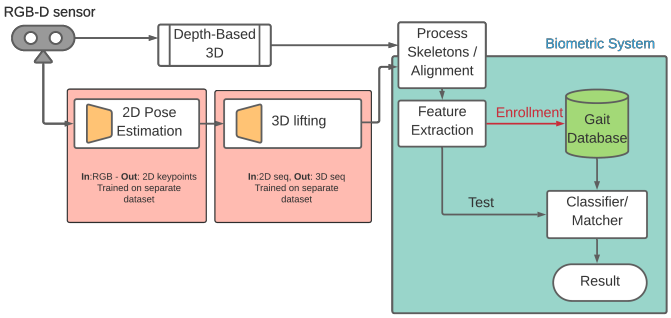

Fig. 5. Overview of our approach. RGB and Depth-image streams are utilized to produce skeleton representations of detected humans. After preprocessing and/or aligning the skeleton sequences, handcrafted Static and Dynamic features derived directly from the positional data are used to carry out Biometric Experiments.

Overall Description We begin with RGB video from a camera source, which we need to decode or to store as frame sequence since every frame will be handled separately. We "encode" every frame into a series of keypoints using 2D pose estimator corresponding to human joints. This can be seen as a form of compression of the useful visual features of the depicted human into data of much lower dimensionality. In that way a lot of covariates that can affect gait (such as lighting, clothing etc) are handled as part of the computer vision front end. The 2D coordinates are then fed into a "lifting" model which tries to infer the corresponding 3D pose.The results of the neural network $F(I, \theta)$ parametrized with θ for image inputs I is the 3D body pose $P = \{p_j\}_{j \in J}$ consisting of 3D locations $p_j = (x_j, y_j, z_j) \in \mathbb{R}^3$ of J body joints with respect to the camera. Since many 3D poses (both anatomically valid or not) can possibly lead to the same 2D projection, the robustness and accuracy of this step is crucial. The 3D poses have to be aligned in a common coordinate system in order to have skeletons that are translated on origin of a coordinate system and have common scale regardless of distance. Rotation has to be handled in a such way that all skeletons are "facing" at the same direction.

Given an aligned sequence of 3D poses human-interpretable features are extracted from the positional joint data. We use those features to perform Gait recognition experiments (see below).

Motion capture ground truth acquisition typically requires expensive marker-suits multicamera systems and/or inertial sensors or other elaborate setups, which were not considered practical in our case. Due to the absence of crucial ground truth data, we opt to work in parallel with a structured light Depth sensor which can provide depth-informed 3D skeleton representations [53]. The depth derived skeletons can serve as a good reference for our pose estimation results.

The 2D and the 3D estimator models are trained separately which allows us some intermediate supervision. A subset

Study	Number and type of features	Description of dynamic features	Feature processing	Recognition	Other Details
Preis et al. [38]	11 static 2 dynamic	Step length, speed	median	1R, C4.5, Naive Bayes	-
Sun et al. [39]	8 static 4 dynamic	Swing angles of leg joints	Discrete Time Warp	Nearest Neighbor	Score level fusion
Ball et al. [40]	6 dynamic	Swing angles of leg joints	Max, mean, Std	K-means clustering	-
Kastaniotis et al. [41]	16 dynamic	8 pairs of Euler angles	Histogram of 40 bins	Gaussian Kernel SVM	-
Choi et al. [42]	4 static 4 dynamic	Position Vectors	-	Linear Matching, Majority voting	Quality adjusted Dissimilarity, Gait cycle phase division

-

Table: Summary of methods using anthropometric skeleton based measurements as features for gait recognition

of the keypoints used in the 2D part are propagated since not all of them are supported by the 3D methods and some of them are not necessary (the subset that we utilize fully can be seen in Figure 6). For joints J connections can be between $J_i = (x_i, y_i, z_i), J_j = (x_j, y_j, z_j)$. The (i, j) is in the set of Φ, and $\Phi = \{(1,2),(2,3),(3,4),(4,5),(3,6),(3,12),(5,9),(5,15),(6,7)$ $,(7,8),(9,10),(10,11),(12,13),(13,14),(15,16),(16,17)\}$.

Additionally the keypoint positions in terms of pixels have to be scaled according to the overall width and height of the image, in order to have a representation that is disentangled from the captured video resolution. For example given a video resolution of 640x480 pixels a keypoint detected at pixel (x,y): 200,200 will be transformed to 2Dposition: 0.3125 ,0.4167

Another in between step that we found necessary was some processing of the 2D coordinates. Depending on the 2D method selected, appropriate interpolation might be needed to fill the gaps of occluded or missed keypoints. We used a simple linear interpolation given valid positions on previous and following frames.

Implementation of pose estimation: Two different models were tried for the **2D pose estimation**, a bottom up and a top-down approach. For the bottom-up approach a tensorflow version of OpenPose [54] was selected as we found the part affinity fields concept to be quite intuitive and also the project has been open sourced which meant we could find a lot of supporting material online. MaskRCNN is top down approach that was also tried to compare performances. An implementation provided with Detectron [55] was used. It seemed to perform better than OpenPose with the added benefit of extracted bounding boxes which turned out to be useful to use on 3d lifting down the line. A major difference of the two was the way that they dealt with occlusions. In OpenPose occluded keypoints were simply not included in the results of a 2D frame. On the contrary occluded keypoints were still included in the MaskRCNN results, they were just placed in a default position. An apparent benefit of the bottom up solution would be the constant inference time for any number of people in the image. A top-down on the other hand performs approximately linearly with the number of people.

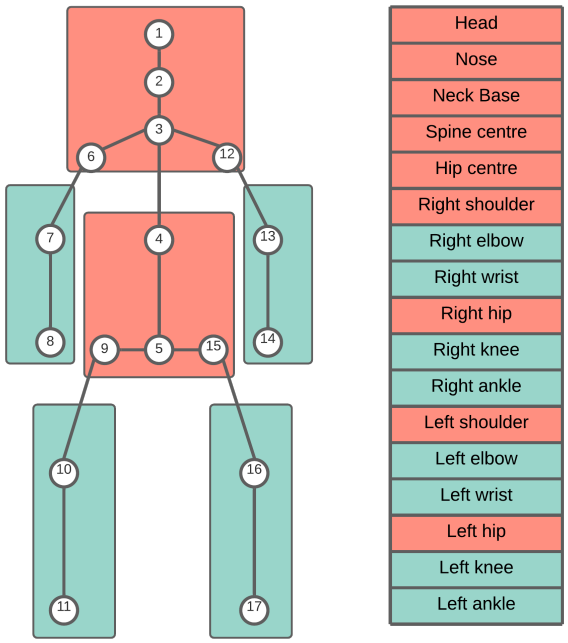

Fig. 6. Human body joint keypoints used to describe pose and motion in this work. It is a selection of points that allows for sufficient description of gait, while acting as a common subset to model different representations. Keypoints in red are upper and lower torso points that exhibit minimal movement during gait. They are used for centroid and direction based skeleton alignment. Keypoints in blue exhibit movement of apparent periodicity and are used for gait modeling.

This difference was not utilized somehow in this work since we chose to work with single person estimation. For both methods, models pretrained on the COCO dataset were used. For the task of **3D pose estimation** two different methods were also tried. Initially we utilize a single-frame "lifting" model [29] to handle each resulting 2D skeleton separately. Preliminary results were not satisfying in a lot of cases as some individual frames whose 2D keypoints were noisy or missing were producing inaccurate results and in some cases were not resembling the actual pose in the slightest. Additionaly

after assembling results of all frames in a sequence the final movement of the 3D skeleton was very noisy and had jitter. To obtain more accurate and smoother results we utilized a model appropriate for handling sequences of 2D poses from video [28]. This method makes use of convolutions in time which lead to very lifelike sequences that resembled the gait closely. The 3D pose methods mentioned were pretrained on Humans3.6M dataset.

2.5D Pose Representation

Since we are working on a monocular basis the exact depth of a persons location in a scene cannot be fully known. For this we adopt a 2.5D pose representation $P^{2.5D} = \{p_j^{2.5D} = (u_j, v_j, z^r{}_j\}_{j \in J}$ where u_j v_j are the 2D projection of the body joint j on a camera plane and $z^r{}_j = z_{root} - z_j$ represents the metric depth with respect to the root joint. As root joint the central pelvis keypoint is used. This decomposition of 3D joint locations into their 2D projection and relative depth is advantageous for in-the-wild images where only 2D pose annotations can be used. However, this representation does not account for ambiguity in scale present in the images, which in turn can lead to some ambiguities in predictions.

Skeleton Sequence Alignment

To have a robust feature extraction we need to align skeleton of all sequences on a common coordinate system, that can apply both to our infered 3D poses and the Kinect depth based ones. Since points close to the center of the body showcase the least amount of movement and irregularity during gait, we employ a similarity transformation on all non aligned skeletons through a centroid point. The centroid can be calculated as the point that is in the middle of the mean positions of upper and lower torso points as seen in Figure 6.

First a translation vector to move the centers of all skeletons to the origin is defined:
$T(t) = p_c(t) = [x_c(t) y_c(t) z_c(t)]^T \in R^3$ where $p_c(t)$ is the centroid of the joint positions in the torso at the t-th frame. Second a scale value to make all skeletons equal in overall size is defined as:
$S(t) = ||p_{uc} - p_{lc}||_2 \in \mathbb{R}$ where p_{uc}, p_{lc} are the centroids of the joint positions of upper and lower torso at t-th frame respectively. Their distance is less dependant on the gait swing compared to distance from head to foot, making the more stable and appropriate for scaling. Thirdly a rotation matrix of three unit vectors is defined:
$R(t) = [\widehat{r}_{mov}(t) \widehat{r}_{left}(t) \widehat{r}_{top}(t)] \in \mathbb{R}^{3x3}$. These unit vectors are denoting the moving, left and top direction and represent the new cartesian coordinate system axis of the aligned sequence. In order to define those vectors the position of centroids will again be used.
In the moving direction: $\widehat{r}_{mov}(t) = \frac{p_c(t) - p_c(t - Tm)}{||p_c(t) - p_c(t - Tm)||_2}$ where Tm is an appropriate time interval for finding moving direction(we use value 1).
In the top direction: $\widehat{r}_{top}(t) = \frac{p_u c(t) - p_l c(t)}{||p_u c(t) - p_l c(t)||_2}$
In the left direction : $\widehat{r}_{left}(t) = \frac{\widehat{r}_{top}(t) \times \widehat{r}_{mov}(t)}{||\widehat{r}_{top}(t) \times \widehat{r}_{mov}(t)||_2}$ since the previous two unit vectors form a 2D plane we can simply produce the orthogonal left direction by utilizing their cross product. We can then recalculate $\widehat{r}_{mov}(t)$ as the cross product of $\widehat{r}_{left}(t)$ and $\widehat{r}_{top}(t)$ in order to have a fully orthogonal coordinate system.

Finally we can transform the original skeletons using:

$$\widetilde{P}(t) = R^{-1}(t)[P(t) - T(t)]/S(t)$$

, where $S(t)$, $T(t)$ and $R(T)$, are the scale value, translation vector and rotation matrix defined previously, at the t-th frame. The position $p(t)$ in the original coordinate system with X, Y and Z axis is converted to the position $\widetilde{p}(t)$ in the new coordinate system with M, L, and T axes. This centroid-based alignment manner helps that all skeletons are well aligned even though a few joint positions are incorrectly estimated.

Spatio-temporal features

We design human interpretable hand crafted features to characterize the gait directly from the positional data. These features or their combination are suitable to encode both the static anthropometric measurements as well as the dynamics of movement.

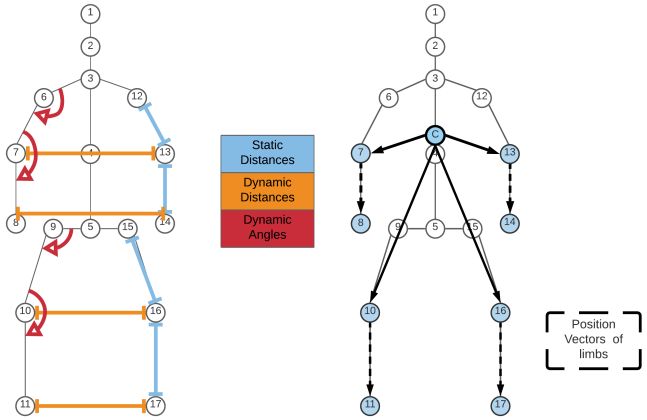

Fig. 7. Types of gait describing features that can be extracted. Left: Static and Dynamic Distance based features, Swing angle features. Right: Centroid based Position vectors as features. Since they are defined with respect to the centroid of a Skeleton aligned sequence, position vectors can encode both the static and dynamic aspect of gait.

Distance features can be defined as $f_{dist}(t) = ||J_i(t) - J_j(t)||_2$ where $i, j \in \widetilde{\Phi}$ and $\widetilde{\Phi}$ can contain all combinations of Joint indexes. When a pair of joint belongs in Φ (consecutive joints) the distance is a **static** distance that represents the limb length. In other pairs due to movement and body articulation the distance will showcase **dynamic behaviour**. For example the index pair (10,11) is a static distance that corresponds to the length of the right shin,which should be constant in time, while the pair (11,17) is the inter-ankle distance, which can be thought of as the stride length and it is a function that changes over time. We make use of 8 static distances and 4 dynamic distances(ankle-to-ankle, knee-to-knee,elbow-to-elbow,wrist-to-wrist).

Angle Features

Joint angles are a type of feature that have been used extensively in previous works to describe the dynamics of

gait,both in 2D and in 3D approaches. In a single frame they are defined as:

$$f_{angle} = \{(\alpha_{ij}, \beta_{ij}) | (i,j) \in \Phi\}$$

$$\alpha_{i,j} = \begin{cases} \arctan \frac{y_i - y_j}{x_i - x_j} & x_i \neq x_j \\ \frac{\pi}{2} & x_i = x_j \end{cases}$$

$$\beta_{i,j} = \begin{cases} \arctan \frac{z_i - z_j}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}} & (x_i-x_j)^2+(y_i-y_j)^2 \neq 0 \\ \frac{\pi}{2} & (x_i-x_j)^2+(y_i-y_j)^2 = 0 \end{cases}$$

Here x, y, z refer to coordinates of our Keypoints2D3D pairs that belong in Φ

When those angles are plotted over time one can examine their dynamic behaviour. We utilize a subset of Φ that produces 8 dynamic angles of limb joints.

Position Vectors as features

Instead of the distance or angle, we can model the spatial walking pattern at each frame as a combination of position vectors, as illustrated on the right side of Fig. 7. In the MLT coordinate system that we defined, a position vector from the joint a to the joint b is expressed as $\overrightarrow{u}_{a \to b}(t)$ Especially, when the starting joint a the centroid, the position vector is represented as $\overrightarrow{u}_{b}(t)$ by omitting a for simplicity. The position vector contains the information about both the distance and angle between the two joints a and b. In other words, it includes both the static and dynamic features of human pose. It is also invariant to view and scale due to the preceding alignment. We construct a frame-level feature vector based on position vectors. Using eight position vectors(4 static and 4 dynamic) we concatenate them to form a 24-dimensional feature vector.

$$f_{vectors}(t) = \begin{bmatrix} \overrightarrow{u}_{L.elbow}(t) \\ \overrightarrow{u}_{R.elbow}(t) \\ \overrightarrow{u}_{L.knee}(t) \\ \overrightarrow{u}_{R.knee}(t) \\ \overrightarrow{u}_{L.elbow \to L.wrist}(t) \\ \overrightarrow{u}_{R.elbow \to R.wrist}(t) \\ \overrightarrow{u}_{L.knee \to L.ankle}(t) \\ \overrightarrow{u}_{R.knee \to R.ankle}(t) \end{bmatrix} \in R^{24}$$

Gait cycle Detection

In order to compare patterns from different sequences it is necessary to extract subsequences that represent a single gait cycle. The most straightforward way to achieve that is to crop the overall sequence at specific points that signify the beginning and ending of a gait cycle, namely when the subject is on single-support of the same foot.

We utilize the inter-ankle distance as a substitute for stride length. It is a also a dynamic signal with double the frequency of the gait cycle. The maxima and minima of this signal showcase a chance in cycle phase, therefore we detect a gait cycle begining when the the distance is at a minimum and ending after 2 more minima (feet are again close together and the same foot is about to begin).

Since the raw positional data might be noisy we use a moving average filter to smooth out the signal and detect maxima and minima more robustly. The figure below illustrates our gait cycle detection.

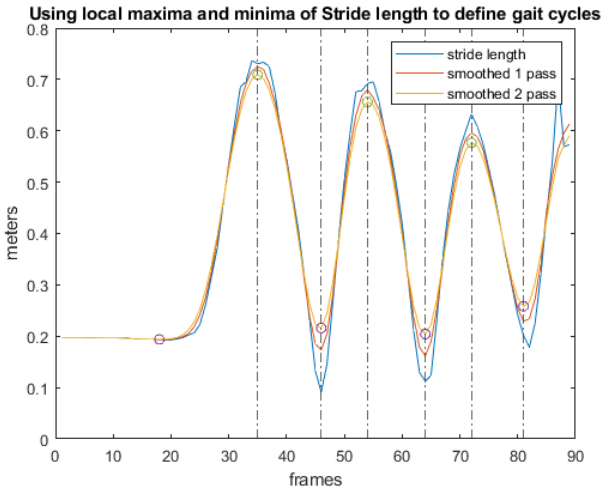

Fig. 8. Gait cycle detected within frames 19 and 63, using a sequence of maxima and minima of smoothed stride length as anchor points

Feature aggregation.

Since we calculate the feature values at every frame in a video it is necessary to aggregate them somehow to obtain features that characterize the whole video and are directly comparable with each other.

For the **static** features this process is straightforward as they are supposed to be constant throughout the video and any fluctuation that exists can be attributed to noise. We use the median value over all the values in the video, as it typically more robust that the mean value that can be adversely affected by noise or poor skeleton detection. For the case of Kinect-depth skeletons we also utilize the depth of the detected subject to characterize frames as being reliable or not, since we know that the extraction can suffer at the extremes of the capturing range. Frames where the person is detected within the interval [1.8m to 3m] from the sensor are considered reliable while the rest are not.

Dynamic template matching: Discrete Time Warping

Dynamic features are represented as time series data. Alignment of such temporal gait data is a challenging task due to variation in walking speed, which might lead to variable length of sequences sequences for the same person. Therefore,applying traditional classifiers in this scenario requires extra pre-processing steps, such as resampling. However, resampling of time-sequence data involves deletion or adding new data, which might affect the recognition performance. On the other hand, non-linear time-sequence alignment techniques can effectively reduce the effect of variable walking speed by warping the time axis. Dynamic time warping (DTW) is a well-known non-linear sequence alignment technique and is utilized for the comparison of our dynamic angles series.

Let θ_{train} and θ_{test} be two sequences of dynamic angles to be compared where the length of θ_{train} and θ_{test} is represented as $|\theta_{train}|$ and $|\theta_{test}|$ respectively.

$\theta_{train} = \alpha_1, \alpha_2, \alpha_3, ...\alpha_{|\theta_{train}|}$
$\theta_{test} = \beta_1, \beta_2, \beta_3, ...\beta_{|\theta_{test}|}$

Here α_t, β_t are the angle values at time t. Given these two time series DTW constructs a warp path $W = w_1, w_2, w_3, w_L$, where $max(|\theta_{train}|, |\theta_{test}|) < L < |\theta_{train}|+|\theta_{test}|$. Here L is the length of the warp path between the two sequences. Each element of the path can be represented as $w_l = (x, y)$ where x and y are indexes of the two compared sequences. A number of constraints must apply to DTW. First the warp path must start at $w_1 = (1, 1)$ and end at $w_L = (|\theta_{train}|, |\theta_{test}|)$. This ensures that every index of both time series is used in the path construction. Second if an index i from $|\theta_{train}|$ is matched with an index j from $|\theta_{test}|$it is prohibited to match any index $> i$ with any index $< j$ and vice versa. This restriction ensures that the warp does not go back in time. Given those restrictions the optimal warp path can be defined as the minimum distance warp path $dist_{optimal}(W)$:

$$dist_{optimal}(W) = min \sum_{l=1}^{L} dist(w_{li}, w_{lj})$$

where w_{li},w_{lj} are two indexes from θ_{train} and θ_{test} respectively and as $dist(w_{li}, w_{lj})$ the Euclidean distance is used.

The basic DTW is extended to a kernel to compute the dissimilarity between a training and a testing gait sample, each of which is a collection of dynamic angles. This kernel aligns the training and testing dynamic sequences of the same angles with each other and computes a match score between them. Summation of all the match scores obtained from the different angle sequences from the training and testing samples is treated as the final dissimilarity measure. The kernel can be defined as:

$$\Delta(\theta, \theta') = \sum_{m=1}^{M} \left\{ min \sum_{l=1}^{L} dist(w_{m,li}, w_{m,lj}) \right\}$$

where $\theta = \theta_1, \theta_2, \theta_3, ..\theta_M$ and $\theta' = \theta_1, \theta_2, \theta_3, ..\theta_M$ are sets/collections of dynamic angles. Different sets of angles lead to different results depending on various factors such as the significance of particular angles in the character of gait, occlusions etc. The set of angles used is of length $M = 8$.

Gait Recognition

For the biometric comparison we perform verification and identification tests using the features described in the previous parts of this section. For the static features recognition is performed in dissimilarity space using the L1 norm to calculate the distance between the aggregated features. For the dynamic features the DTW kernel score is used as the dissimilarity measure. Since different types of features carry different information, recognition results are fused on rank level. During experiments we will refer to this as **Method1**. For vector based features recognition is performed on frame level similarly with the approach in [42]. To aggregate results of multiple frame-level decisions a simple majority vote is adopted. During experiments we will refer to this as **Method2**.

IV. Experimental Setup

To test our overall methodology we require videos of single-person gait sequences. In order to have a basis of comparison synchronized depth stream is required to produce depth-informed 3D skeletons that can be used as reference and directly compared to the ones inferred by our image-2D-3D method. Since the two types of skeletons differ vastly, we need to perform biometric experiments on each modality separately.

The robustness of the methods needs to be tested with the following covariates and conditions in mind:

- Viewing angle. If we are to prove view-invariance the data collected have to produce promising results regardless of the viewing angle that a sequence is captured from.
- Uniqueness of features. The features used have to be sufficiently discriminative, especially considering a small number of participants.
- Permanence. Gait behaviour might change over time so we need to acquire data from different recording sessions. This way we implicitly test for clothing invariance (somewhat) as well.
- Gait style, parallel activity. Gait is not always performed in its purest form but it is usually done alongside other small actions or activities. The data collected shall somewhat account for that.

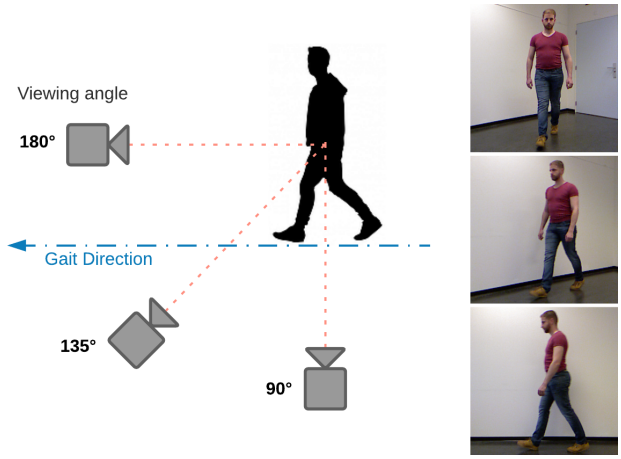

Fig. 9. View angle variants in the acquired dataset. Gait sequences were captures with camera positioned at frontal, oblique and vertical angles to the direction of motion.

A KinectXBOX360 depth sensor was available to us and used for the experiments. We utilize the RGB video stream and Depth stream which were captured in a synchronized manner. A custom C# application that connects to the Kinect-for-Windows SDK was built to capture and save both streams on memory using buffers upon the frame grabber. For the framerate of 30fps the maximum possible resolution provided by the Kinect is 640 x 480 pixels for RGB and 320x240 for the Depth stream.

Three camera positions were selected to test for view-invariance. A lateral view, an oblique, and a frontal view cover the three basic cases of gait capture as the Figure

Fig. 10. Sample frames from a male(top) and female(bottom) subject of our acquired Twente Gait Dataset. Walking conditions from left to right: normal gait, carrying, interacting with phone, waving/signaling, sudden mid-gait stop and start

suggests. The RGB-D sensor was positioned on a tripod and elevated at about waist level $\approx$ 85cm from ground. Gait sequences were performed in such a way that the maximum Operative distance of the Depth sensor is utilized. Subjects were requested to perform different actions alongside some of their sequences to try and cover a bigger space of valid walks.

Dataset acquisition

A short break-down of the recorded data:

- 10 volunteering participants as subjects (5M+5F)
- 2 recording sessions per participant
- 3 camera positions, viewing angles, separate gait sequences captured sequentially
- 5 walking "types".
  1) Normal walking
  2) Normal walking with carrying (backpack)
  3) Normal walking with passive hands(action: talking/interacting with phone)
  4) Normal walking with active hands(action: greeting/waving/signaling)
  5) Sudden stop and start while walking
- 2 sequences per walking type

This amounts to 600 sequences (60 per subject) to be used for various recognition experiments, using different protocols (gallery and probe sets).

Verification/Identification tests

A comparison of all entries in the dataset with each other produces a complete score matrix. By dividing the entries of this matrix into Genuine and Imposter scores we can examine the performance of the system in a verification manner. Verification experiments are done both for the static and the dynamic feature separately.

We perform Identification experiments using the two types of skeleton datasets (Kinect , Vpose) that we recorded, and two methods of features and processing (Method1, Method2). In the **overall experiment** we use random permutations of the dataset which is split in half into equally sized Training(gallery) and Testing(probe) sets, in order to gauge overall performance. After multiple randomized splits (50 iterations) the results are averaged for each modality/method. We report correct identification rate at rank1 to rank5.

Three more experiments are defined to measure identification performance using different splits of the datasets. Correct Identification Rate (rank-1 accuracy) metric is used for comparison.

Split 1: This is again a split in half (30 videos for training and 30 for testing for each subject) but this time the sets are separated based on the recording session i,e, videos captured during session 1 are used for training and videos from session 2 are used for testing and vice versa. This is done to test the permanence of gait features, since it is expected that samples from the same recording session can be quite correlated to one another.

Split 2: In this test we form the entries into gallery and probe sets based on viewing angle. The dataset is divided into the 3 views (A1:lateral, A2:oblique, A3:frontal). One third is used as gallery and the other two are concatenated into the probe. This is expected to give us further insight on the view-variance/invariance question.

Split 3: A small and final experiment splits the dataset into 5 parts based on walking "types" and utilizes only the first type (normal walking) as seen data. So for each subject 12 gait videos are used as gallery and 48 as probe. This test will show whether 'imperfect' gait samples can be successfully matched to fewer but more 'proper' gait patterns.

V. Experimental Results and Analysis

The quality of source data and acquisition methods are crucial to produce results that can be used for biometric recognition. Since we approach the problem with view-invariance in mind, it is important to know how well the methods perform on each of the three captured views. One way to quantify that is to measure the number of missed joints, that is keypoints that were not detected in the pose estimation part. Those missed keypoints typically signify occlusions or inaccuracies of the method on a given frame. Table below illustrates the range of missed keypoints on sequences of different subjects grouped by view.

View	Missed keypoints(%)
Frontal	[4-6]%
Oblique	[7-9]%
Lateral	[11-18]%

We notice a significant number of missed keypoints on lateral view.This is as expected since subjects are seen from the side leading to a large portion of the body, swinging limbs like the back facing arm and leg being occluded in multiple frames. This percentage could be in fact larger but we see that the pose estimation models are able to infer positions of some ambiguous joints. Missed points of frontal view can be attributed to lower scale of appearance of the subject at distance and partly because of keypoints that are out of the field of view of the camera (The subjects head and feet can exit the frame at the later stages of a sequence). In the oblique view case both types of inaccuracies are present but to a lower degree. An analysis of most occluded keypoints, grouped by view can be found in the Appendices section.

Results of alignment

Using the alignment method described on section 3, we obtain a ”point-cloud” of keypoints that are centered around the torso centroid for every sequence. This process is done both for Depth and RGB based skeletons and the two are then overlapped for visual comparison. We notice some strong biases to develop over time that hints that our 3D pose method is not totally view-invariant.

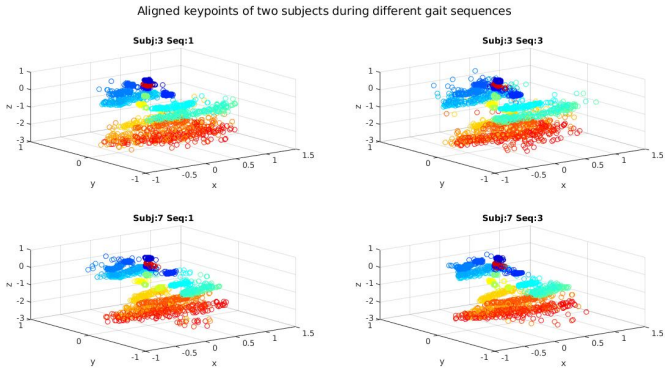

Fig. 11. Aligned gait sequences based on torso centroid from two different subjects. Each joint type is plotted with a different color for clarity. Notice that beside the spurious data the two sequences of each subject maintain some characteristic attributes.

Extraction of features

We extract static and dynamic features for skeletons produced from all sequences in our dataset using both Depth and RGB modalities.

Static features extracted from Kinect seem to be more consistent and have better inter-class separability. Even though the intitial positional signals may be quite noisy, the presence of depth knowledge leads to distances that after some statistic aggregation point to specific subjects. They seem to be valid for anthropometric measurements. On the other hand dynamic features might not be as reliable due to the way that the Kinect skeletons are constructed. For example there are no pose restrictions regarding anatomical plausibility with might lead to effects such as knees bending backwards , momentary leg swapping etc.

Features extracted from RGB-based skeletons are riddled with different problems. Static features appear more stable, less noisy, within sequences, but inter class separation seems to be reduced as well. That means that every skeleton is not totally representative of the subject but is somewhat averaged towards a common representation. This can be possibly explained as the effect of depth ambiguity that is inherent to our approach and the way that the 3D positions of joints are inferred. It means that the method is not totally ”identity preserving”.

Gait recognition results

We report verification results using the static features. Figure 14 shows the resulting ROC curve and equal error rate point plotted on top. We notice that with the use of only reliable frames Kinect skeletons are able to be successfully identified based on their static features with a high probability. On the other hand RGB keypoint based skeletons somewhat lose this ability partly due to depth ambiguity and insufficient generalization. The identity information seems to be heavily filtered by the method. Results of verification test using dynamic features (Figure 15) lead to an Equal Error Rate of approx. 0.253, which shows that dynamic features maintain some of their discriminating power. Although examination of histogram of impostor vs genuine scores shows that there is significant overlap between the scores, which hinders the recognition .

Identification results for the overall experiment can be examined in Figure 16. It can be confirmed again from these results that Kinect based skeletons retain more identifying information than inferred ones, despite being seemingly more noisy. The frame-level matching approach (Method 2) using position vectors as features seems to outperform our DTW method based on dynamic features. This might hint that position vectors are a better representation or that decision at frame-level even though it is very costly can be more robust to irregularities in gait behaviour.

Identification results (rank1 accuracy) for the first split can be seen in the next table. The rate of identification seems generally lower than the respective methods on the randomly split test. This signifies that there might be strong correlation between consecutively recorded gait samples from the same enrollment session.

Kinect	Train1Test2	Train2Test1
Method1	67.33%	73.67%
Method2	63.33%	74%

Vpose	Train1Test2	Train2Test1
Method1	73.67%	67.67%
Method2	86.67%	84.67%

Table:Identification Results (CIR) for Split1 test

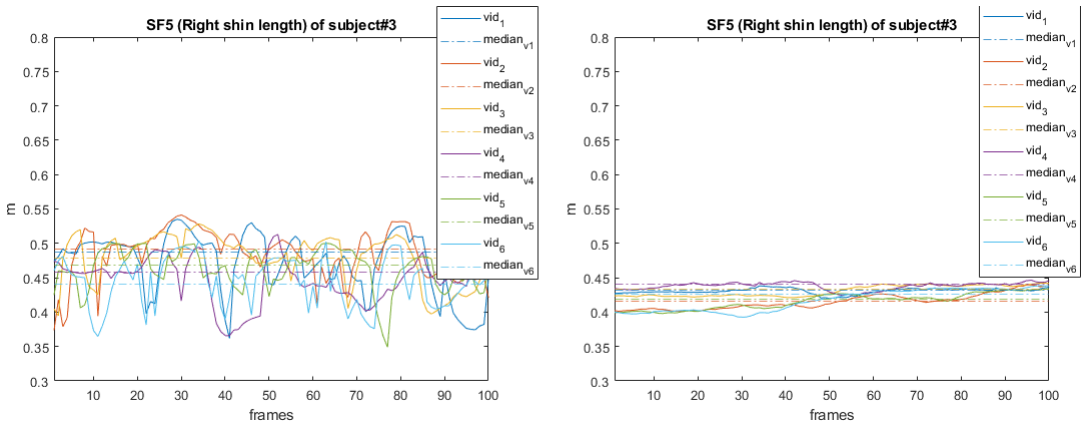

Fig. 12. Static feature extraction, from Kinect skeletons(left) and Keypoints2D3D(right). Single static feature(R.shin length) from 2 gait sequences for each viewing angle.

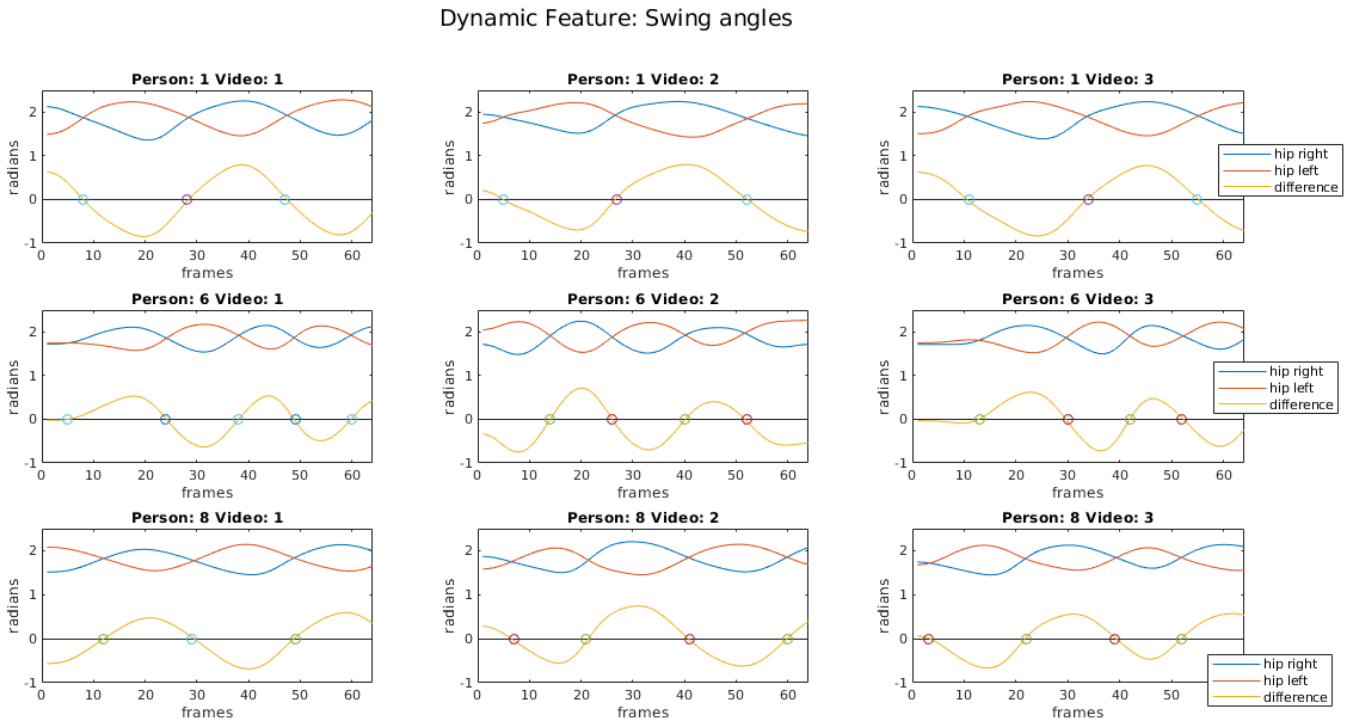

Fig. 13. Example of a dynamic feature:The swing angle of the legs at hip level. Three different sequences from 3 different subjects

The next experiment (based on Split 2) tests for view invariance.

Kinect	A1vsA2A3	A2vsA1A3	A3vsA1A2
Method1	67.75%	66.75%	68.5%
Method2	45%	60%	58%

Vpose	A1vsA2A3	A2vsA1A3	A3vsA1A2
Method1	30%	27%	16%
Method2	50%	70%	46%

Table:Identification Results (CIR) for Split2 test. (View angles: A1-lateral, A2-oblique, A3-frontal).

Results suggest that our approach does not lead to view-invariant recognition. Using dynamic angle features seems to lead to very poor results, especially in the case where the frontal view is the only data in the gallery. Kinect skeletons seem to be more robust to view-angle change. It is even interesting to see that frontal view from Kinect leads to some of the best results (possibly because frontal view of the subject was how the Kinect sensor was meant to be used).

The last experiment is based on Split 3. In this case the gallery consists only of normal walking entries and the more

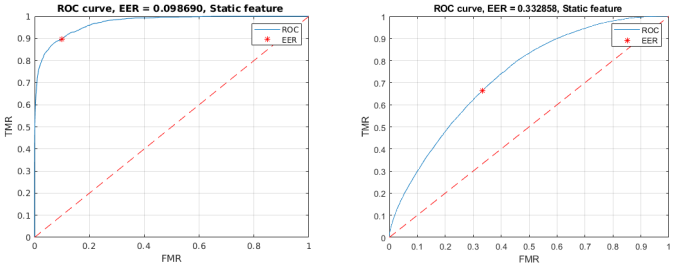

Fig. 14. Verification results:ROC curve produced for a moving threshold of static feature similarity.Kinect(Left),VPose(Right)

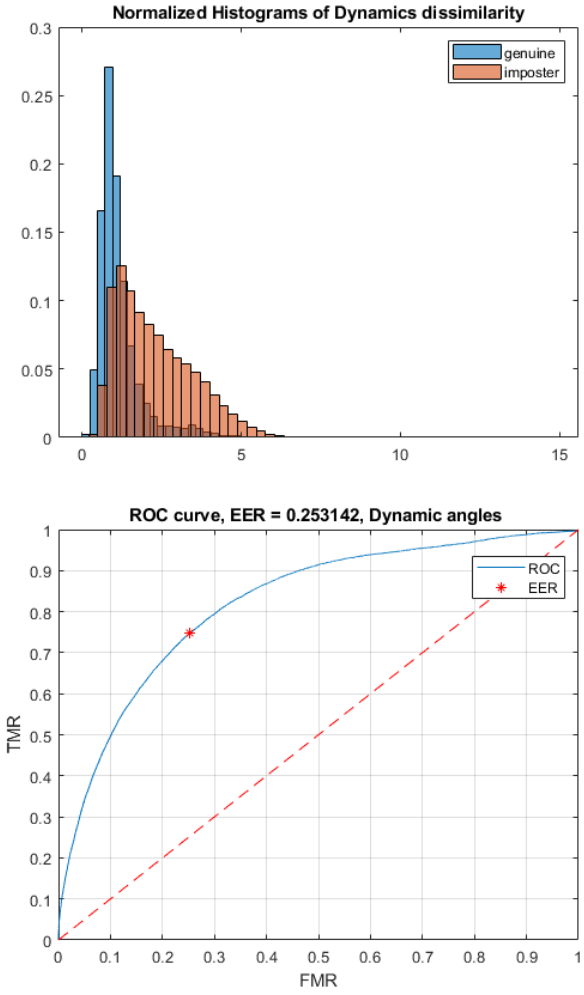

Fig. 15. Verification results: System performance using Dynamic features. Genuine and Imposter dissimilarity scores histogram(top). ROC curve for Vpose(bottom)

difficult types of gait are used as probe.

	Kinect	Vpose
Method1	80%	72.29%
Method2	73.33%	77.71%

Table:Identification Results (CIR) for Split3 test

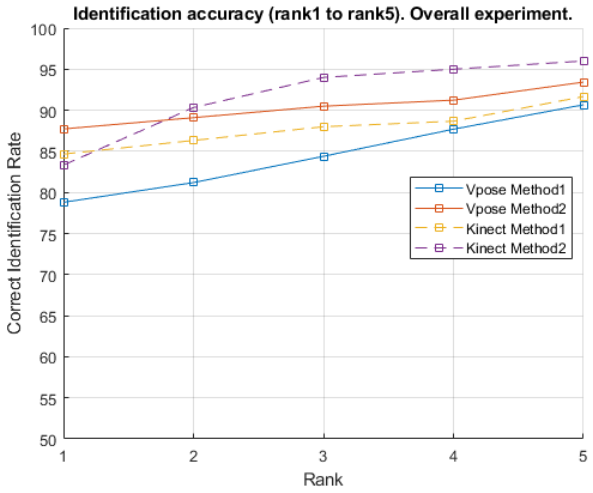

Fig. 16. Identification overall results

The results point out that recognition is possible even given more complex gait types, but one should try to obtain a more diverse gallery of samples when enrolling a subject for the purposes of gait recognition.

VI. Conclusion

The goal of this work was to work towards the realization of a view-invariant gait recognition method using RGB video from a single camera. It proved to be a challenging task with multiple subproblems and points requiring attention. We rely on recent human pose estimation works to initially extract 2D keypoints from RGB frames and then try to infer the 3rd dimension component from them. A robust alignment method is used to transpose all skeleton representations in a way that they become directly comparable. Human interpretable hand-crafted features are extracted to provide the basis for gait based recognition. A small dataset of volunteering subjects is collected to assess the validity of our methodology. Results are compared to respective ones produced using a Depth sensor. We report results that show potential about the realization of a recognition system with the required properties. A rank-1 accuracy of up to 88% is achieved on the overall experiment. The dataset we use for validation is rather small to help us reach definitive answers but some phenomena can be observed. We noticed that the model, which we use to reach a 3dimensional representation of human pose, is possibly over-filtering/smoothing the identity information in our data. The method is not free of view angle dependencies since some biases can occur. Lastly, the simple features we used to describe gait, might not be adequate since they do not offer sufficient/satisfactory separation of subjects. Based on the above we propose some aspects that future work on the topic should focus on. A dataset of longer gait sequences is needed if one wants to explore and exploit the periodicity of gait related signals. Higher level or learnable features could potentially offer a more complex and complete description

of walking patterns. Finally, different methods of 3D pose lifting (potentially anatomy and occlusion aware ones) should be investigated, in the search for a more suitable model.

REFERENCES

[1] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis, "Gait recognition: a challenging signal processing technology for biometric identification," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 78–90, 2005.

[2] L. Sloman, M. Berridge, S. Homatidis, D. Hunter, and T. Duck, "Gait patterns of depressed patients and normal subjects." *The American journal of psychiatry*, 1982.

[3] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," 2020.

[4] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," 2019.

[5] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, pp. 55–79, 01 2005.

[6] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1014–1021.

[7] M. Burenius, J. Sullivan, and S. Carlsson, "3d pictorial structures for multiple view articulated pose estimation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3618–3625.

[8] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 2878–90, 12 2013.

[9] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2014.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, p. 2012.

[11] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[12] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," 2015.

[13] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," 2016.

[14] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[15] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," 2018.

[16] [Online]. Available: https://github.com/MVIG-SJTU/AlphaPose

[17] [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[18] [Online]. Available: https://cocodataset.org/#keypoints-leaderboard

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018.

[20] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," 2018.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[22] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[23] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," 2017.

[24] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3d human pose from structure and motion," 2018.

[25] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," 2018.

[26] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[27] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[28] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," 2019.

[29] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.603

[30] R. Zhao, Y. Wang, and A. Martinez, "A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image," 2016.

[31] C.-H. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," 2017.

[32] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," 2017.

[33] K. Lee, I. Lee, and S. Lee, "Propagating lstm: 3d pose estimation based on joint interdependency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[34] X. Chen, K.-Y. Lin, W. Liu, C. Qian, X. Wang, and L. Lin, "Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation," 2019.

[35] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3d human pose estimation," *Lecture Notes in Computer Science*, p. 69–86, 2018.

[36] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, "Anatomy-aware 3d human pose estimation with bone-based pose decomposition," 2021.

[37] Y. Cheng, B. Yang, B. Wang, Y. Wending, and R. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 723–732.

[38] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien, "Gait recognition with kinect," in *1st international workshop on kinect in pervasive computing*. New Castle, UK, 2012, pp. 1–4.

[39] J. Sun, Y. Wang, J. Li, W. Wan, D. Cheng, and H. Zhang, "View-invariant gait recognition based on kinect skeleton feature," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 24 909–24 935, 2018.

[40] A. Ball, D. Rye, F. Ramos, and M. Velonaki, "Unsupervised clustering of people from'skeleton'data," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 225–226.

[41] D. Kastaniotis, I. Theodorakopoulos, G. Economou, and S. Fotopoulos, "Gait-based gender recognition using pose information for real time applications," in *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–6.

[42] S. Choi, J. Kim, W. Kim, and C. Kim, "Skeleton-based gait recognition via robust frame-level matching," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2577–2592, 2019.

[43] X. Wu, S. Yu, and Y. Huang, *Multiscale Temporal Network for Video-Based Gait Recognition*, 10 2019, pp. 75–83.

[44] I. Rida, A. Bouridane, G. L. Marcialis, and P. Tuveri, "Improved human gait recognition," in *Image Analysis and Processing — ICIAP 2015*, V. Murino and E. Puppo, Eds., 2015.

[45] H. Guo, B. Li, Y. Zhang, Y. Zhang, W. Li, F. Qiao, X. Rong, and S. Zhou, "Gait recognition based on the feature extraction of gabor filter and linear discriminant analysis and improved local coupled extreme learning machine," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–9, 04 2020.

[46] C. Song, Y. Huang, Y. Huang, N. Jia, and L. Wang, "Gaitnet: An end-to-end network for gait based human identification," *Pattern Recognition*, vol. 96, p. 106988, 2019.

[47] M. Balazia and P. Sojka, "Gait recognition from motion capture data," 2017.

[48] M. Ahmed, N. Al-Jawad, and A. Sabir, "Gait recognition based on kinect sensor," vol. 9139, 05 2014, p. 91390B.

[49] Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 149–158, 2004.

[50] T. Krzeszowski, B. Kwolek, A. Michalczuk, A. Świtoński, and H. Josiński, "View independent human gait recognition using markerless 3d human motion capture," in *Computer Vision and Graphics*, L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, and K. Wojciechowski, Eds., 2012.

[51] N. Li, X. Zhao, and C. Ma, "Jointsgait:a model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping," 2020.

[52] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020.
[53] 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 *CVPR 2011*, 2011, pp. 1297–1304.
[54] [Online]. Available: https://github.com/infocom-tpo/tf-openpose
[55] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron," https://github.com/facebookresearch/detectron, 2018.

APPENDIX

Occlusion Analysis. Most occluded joint keypoints grouped per viewing angle.

[1]

Side view	Body Joint	Occluded frames
1st Most occluded	L.elbow	81 %
2nd m.o.	L.shoulder	77 %
3rd m.o.	L.wrist	64%
4th m.o.	R.hip	17%
5th m.o.	L.knee	11%

[2]

Oblique view	Body Joint	Occluded frames
1st Most occluded	L.elbow	51 %
2nd m.o.	L.shoulder	42 %
3rd m.o.	L.wrist	28%
4th m.o.	R.hip	14%
5th m.o.	L.knee	8%

[3]

Frontal view	Body Joint	Occluded frames
1st Most occluded	L.hip	17 %
2nd m.o.	L.ankle	16 %
3rd m.o.	Head	11%
4th m.o.	R.elbow	5%
5th m.o.	L.wrist	4%