Chasing feet in the wild: A proposed egocentric motion-aware gait assessment tool

Mina Nouredanesh[1,2][0000−0002−5768−0348], Aaron W. Li[2,3], Alan Godfrey[4][0000−0003−4049−9291], Jesse Hoey[2,3][0000−0001−5340−2204], and James Tung[1,2][0000−0002−0771−2313]

[1] Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, Canada {m2noured,james.tung}@uwaterloo.ca
[2] AGE-WELL NCE Inc (Canada's technology and aging network), Toronto, Canada
[3] David R. Cheriton School of Computer Science, University of Waterloo, Canada w89li@edu.uwaterloo.ca, jhoey@cs.uwaterloo.ca
[4] Department of Computer and Information Science, Northumbria University, Newcastle upon Tyne, UK alan.godfrey@northumbria.ac.uk

Abstract. Despite advances in gait analysis tools, including optical motion capture and wireless electrophysiology, our understanding of human mobility is largely limited to controlled conditions in a clinic and/or laboratory. In order to examine human mobility under natural conditions, or the 'wild', this paper presents a novel markerless model to obtain gait patterns by localizing feet in the egocentric video data. Based on a belt-mounted camera feed, the proposed hybrid FootChaser model consists of: 1) the FootRegionProposer, a ConvNet that proposes regions with high probability of containing feet in RGB frames (global appearance of feet), and 2) LocomoNet, which is sensitive to the periodic gait patterns, and further examines the temporal content in the stacks of optical flow corresponding to the proposed region. The LocomoNet significantly boosted the overall model's result by filtering out the false positives proposed by the FootRegionProposer. This work advances our long-term objective to develop novel markerless models to extract spatiotemporal gait parameters, particularly step width, to complement existing inertial measurement unit (IMU) based methods.

Keywords: Ambulatory gait analysis · wearable sensors · deep convolutional neural networks · egocentric vision · optical flow

1 Introduction

The lack of clinical information on a day-to-day basis hinders our understanding of disease trajectories on multiple time scales, including diseases affecting gait and balance (e.g., neurological conditions). Free-living (habitual) ambulatory gait analysis has demonstrated unique insight into disease progression, with implications for diagnosis and evaluating treatment efficacy. For example, spatial metrics (e.g., step length), temporal metrics (e.g., step time), and gait irregularities (e.g., compensatory balance reactions or near-falls) of free-living mobility

behaviour have demonstrated promising capabilities to predict the risk of falling in older adult populations.

The recent explosion of ambient sensors (e.g., motion capture sensors, force mats), smart phones, and wearable sensor systems (e.g., inertial measurement units, IMUs) have facilitated the emergence of new techniques to monitor gait and balance control in natural environments and during everyday activities [14, 13, 30]. Embedded into living environments, ambient third-person video (TPV) and depth cameras (e.g., Microsoft Kinect) have been investigated as means to extract gait parameters [36, 37], detect episodes of freezing of gait in Parkinson's disease [34], detect falls, and longitudinal changes in the patient's mobility patterns [35, 33, 38]. While TPV systems have demonstrated potential to detect small changes over long periods (i.e., months to years), these approaches suffer from visual occlusions (e.g., furniture), difficulty handling multiple residents, and extraction of spatiotemporal parameters when the full-body view is unavailable. Moreover, they are restricted to fixed areas. Considering mobility is characterized by moving the body from one location (i.e., environment) to another, significant daily-life mobility data may go uncaptured without multiple camera coverage using ambient sensors.

An alternative approach is to use wearables sensors affixed to the user's body. There have been many successful research programs using IMUs to monitor physical (and sedentary) activity, identify activity types, estimate full body pose, and measure gait parameters [30, 13, 14, 10, 52]. In particular, body-worn IMUs have demonstrated excellent capabilities to measure temporal gait parameters. However, a critical drawback associated with the use of IMUs is inaccurate estimation of key spatial parameters. In particular, step width is linked to gait stability and have a strong association to fall risk [49, 7]. This measurement limitation is largely attributed to a relative lack of motion in the frontal plane during gait, resulting in small IMU excitation and low signal-to-noise ratio.

Egocentric first-person video (FPV), acquired via body-worn cameras, may outperform IMUs for the purpose of estimating spatial parameters of gait. Bearing in mind a waist-worn camera pointed down and ahead of the user, FPV offers a potentially stronger signal for spatial estimation, especially in the frontal plane. There are also secondary reasons for investigating FPV as a sensing modality. Vision captures rich information on the properties of the environment that influence mobility behaviour, including slope changes (e.g., stairs, curbs, ramps) and surfaces (e.g., gravel, grass, concrete) [11, 12]. Furthermore, FPV offers the potential to reconstruct events by capturing the immediate environmental context more readily than IMU-based data alone. Without detailed information of the mobility context, such as the presence of other pedestrians, terrain characteristics, and obstacles, the ability to interpret ambulatory gait data is constrained. For example, FPV recordings have been used for the purpose of validation of other IMU-based algorithms [9, 10] by manually viewing video frames and identifying specific events.

To address the problem of ambulatory measurement of spatial gait parameters, this paper tackles the initial problem of localizing feet in FPV frames

in 2D coordinates of video captured from a belt-mounted camera. We propose a method to generate pixel-wise foot placement outputs towards the eventual goal of estimating spatial parameters (e.g., step width). The transformation between pixel outputs to distances, likely using 2D metrology approaches, is beyond the scope of the current study and will be examined in subsequent works. To achieve the foot localization solution, we first propose a FPV-based deep hybrid architecture called the FootChaser model (see Fig. 3). Comprised of a) the FootRegionProposer model, which uses a ConvNet to propose high confidence feet regions (or bounding boxes), and b) the LocomoNet, which examines the temporal dynamics of the proposed regions to refine the FootRegionProposer output by filtering the false positives to locate feet. An evaluation of the proposed method to accurately localize feet is reported and discussed. Finally, as the problem is new, a FPV dataset (see Fig. 4) is going to be prepared in the near future for benchmark testing of anticipated advancements.

1.1 Related work

While there have been TPV-based research efforts utilizing smart phone or ambient camera video to assess gait (e.g., [33, 36, 37]) and estimate pose (e.g., [40, 39, 16, 41, 19]), the challenges and signals associated with FPV are distinct. There are several factors that challenge the proposed concept: 1) occlusion or extreme illumination conditions, 2) similar objects/terrain patterns to the feet (e.g., other people's feet), and 3) motion blur from fast movements. In this section, we focus on reviewing previous efforts using FPV to address these challenges and to inform our chosen methodologies, i.e. camera type and wear location.

There are relatively few previous works aiming to extract spatial gait parameters using FPV. An interesting and novel approach was using a walker-mounted depth and/or color camera to estimate 3D pose of lower limbs, mainly in frontal plane [27, 25, 26]. To achieve this, Ng et al. [26] used general appearance model (texture and colour cues) within a Bayesian probabilistic framework. In [25], a Kinect (depth) sensor along with two RGB cameras were placed on a moving walker, and the 3D pose was formulated as a particle filtering problem with a hidden Markov model. The key limitation of these works is the dependency on a stable platform (i.e., walker) to afford consistent views of the lower limbs and monitor pose over time, which is not generalizable to individuals that do not require a walking aid for ambulations.

The possibility of using one or several body-mounted cameras is investigated for 3D full body [31, 15, 28] and upper limb (arms and hands) [23, 24] pose estimation. In [31, 15], outward-looking body-mounted cameras along with optimization approaches were used to estimate 3D body pose. In [31] more than ten cameras were attached to all the person's joints, and structure from motion approach was used to localize the cameras, estimate the joint angles and reconstruct human motion. The main limitation of the proposed method is the obtrusive multi-camera setup and intensive computational load required to infer pose in a video sequence. To alleviate the main weaknesses of [31], Jiang et al. [15] developed a model based on synchronized egocentric videos captured by a

chest-mounted camera and a Kinect sensor. The 3D body pose model employs camera egomotion and contextual cues to infer body pose, without direct views of the key body parts (i.e., legs, feet) desired for gait assessment. Moreover, the videos were restricted to relatively static activities (i.e., sitting, standing). Such restrictions and the failure to examine more complex (i.e., dynamic) scenarios limits the applicability is the important limitation of of their approach to the gait assessment problem.

In contrast to the previous studies, [29] and [28] utilized body-related visual cues (outside-in/top-down view) provided by fisheye cameras attached to a bike helmet and baseball cap, respectively. In [28], a ConvNet for 3D body pose estimation was developed to address limitations in its former version [29], including dependency on 3D actor model initialization and inability to run in real-time. Although the authors compensated for the distortion imposed by the fisheye lens, estimation of the lower body 2D heatmaps (ankles, knees, hip, and toes) was less accurate due to the strong perspective distortion (i.e., a large upper body and small lower body).

The closest approach in spirit to the proposed approach is a hybrid method which combines both global object appearance (spatial network) and motion patterns (temporal network) in a two-stream ConvNets structure. This approach was inspired by Simonyan and Zisserman [5], in which a ConvNet was trained by stacks of optical flow for the task of TPV-based activity recognition. Similar architecture is also employed in FPV-based methods to recognize different activities [1, 4]. To capture long-term sequential information from FPV data, recurrent neural network/long-short term memory (LSTM) was used by Abebe et al. [2, 3] where stacked spectrograms generated over temporal windows from mean grid-optical-flow vectors were used to represent motion [4].

Modeling temporal information in a specific regions enclosed by bounding boxes in consecutive frames is investigated in some TPV-based studies [18, 22]. In [21] an object-centric motion compensation scheme was implemented by training CNNs as regressors to estimate the shift of the person from the center of the bounding box. These shifts were further applied to the image stack (a rectified spatiotemporal volume) so that the subject remains centered. More related to our LocomoNet approach is the work by Brattoli et al. [18], in which a fully connected network was trained to analyze the grasping behavior of rats over time. Based on optical flow data of both initial positives (paw regions) and random negatives cropped from other regions, temporal representation was learned to detect moving paws.

2 The FootChaser framework

In this section, we describe the framework for proposing high confidence regions by incorporating both temporal and spatial data, for the task of gait assessment. As an alternative to inferring gait parameters from 3D pose estimates, we hypothesized that tracking the centers of the person's feet in 2D plane of walking over time could provide accurate spatial estimates. The scope of this paper is to

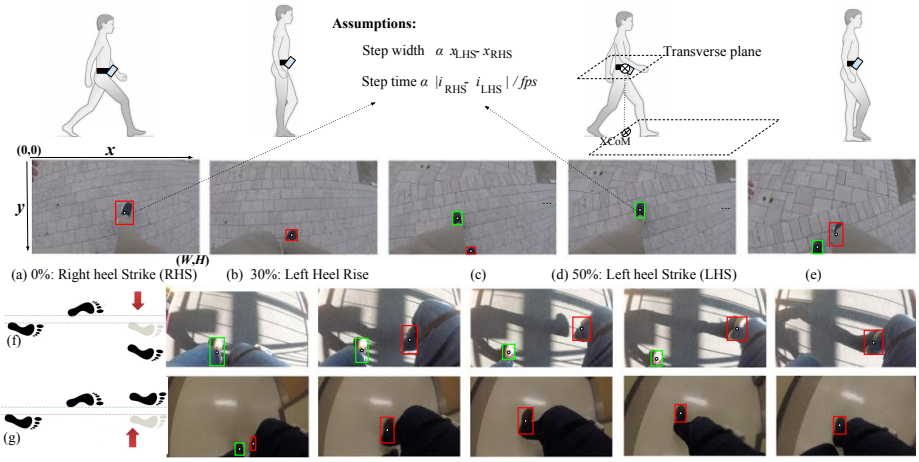


Fig. 1. Egocentric camera-based gait assessment overview. Panels a,b,c,d,e represent different phases of gait captured by a belt-mounted camera. The x and y location of the right foot (red bounding boxes) and left foot (green boxes) over consecutive frames (XCoM: extrapolated center of mass). Rows f and g depict lateral sidestep and lateral crossover compensatory balance reactions, respectively. These reactions are important behaviours related to fall risk. Note the transformation between pixel-wise box coordinates to distances is not covered in the current study.

first detect the feet, and examine the transformation between camera coordinates to spatial locations in subsequent efforts.

Let I_i be the i^{th} frame in a video sequence with the length N, captured by a belt-mounted camera with an outside-in top-down view ($i = \{1, 2 \cdots N\}$). The manually annotated ground truth (GT) data is in the form of bounding boxes $GT_{f,i} = [x^{GT}_{f,i}, y^{GT}_{f,i}, w^{GT}_{f,i}, h^{GT}_{f,i}]$ indicating the camera wearer's feet ($f = \{left, right\}$) in 2D 1080×1920 coordinate system of each frame (see Fig. 1). x and y denote the center ($C^{GT}_{f,i}$), and w and h represent the width and height of the bounding box(es) respectively (see Fig. 2). The goal of the *FootChaser* framework is to detect and localize the centers of each foot (if present in the frame) in the form $P_{f,i} = [x^{P}_{f,i}, y^{P}_{f,i}, w^{P}_{f,i}, h^{P}_{f,i}]$ during the gait. In an ideal case, the error measure (E) will be minimized for the x ($E(x^{GT}_{f,i}, x^{P}_{f,i})$), y ($E(y^{GT}_{f,i}, y^{P}_{f,i})$) trajectories and the underlying area should be the same for the Ps and GTs. The intersection over union (IoU) measure will be maximized ($IoU = 1$). The predicted x ($\approx$ frontal axis) and y ($\approx$ sagittal axis) trajectories can be used to estimate pixel-wise step width and step length gait parameters, respectively.

To investigate the feasibility of pixel-wise step-by-step gait parameter extraction, the x^{GT}_{left}, x^{GT}_{right} data are plotted in Fig. 2. While y^{GT}_{left} and y^{GT}_{right} were examined for measurement of step length, we focus on step width estimation in the current study. We observed that **1)** the trajectories roughly resemble the center of pressure (CoP) data captured by forceplates, **2)** the local maxima and minimuma are correlated with right heel strike (RHSs) and left heel strike

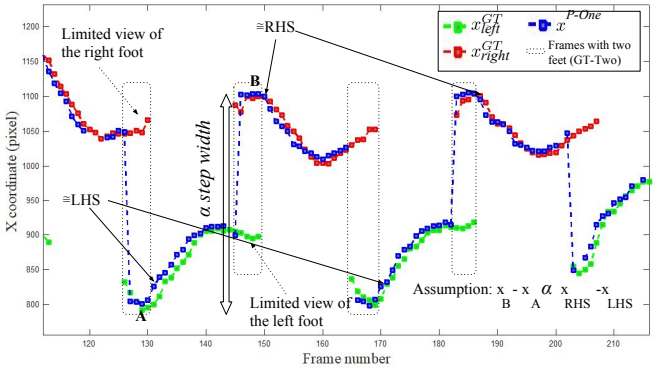


Fig. 2. Sample bounding box X-coordinate time series data from dataset 2. Ground Truth (GT) data for left (green) and right (red) feet, and FootChaser predictions with 1 identified region (blue). Annotated x location of left heel strike (LHS) and right heel strike (RHS) are marked. Periods with 2 identified feet (GT-Two) are indicated by dotted boxes.

(LHSs), respectively, and **3)** GT data can be divided into frames with one foot ($GT-One$), and both feet ($GT-Two$).

In most of the $GT-Two$ frames, a small portion of the trailing foot is observable (see Fig. 1), and is irrelevant for extraction of gait parameters. Considering shape distortions affect detection results, we hypothesized that the ConvNet is more likely to detect the other foot rather than the less-visible one similar to the findings of Huang et al. [45] and Rozamtsev et al. [20]. In other words, in the frames with two GT, the network often locates the center of the foot that is required for the extraction of gait parameters.

Considering these cues, we surmised that tracking each foot separately is unnecessary and frames with only one predicted center (i.e., foot) can be used to extract step width. Specifically, (C_i^{P-one}) obtained from the FootChaser ($P-One = [x_i^{P-one}, y_i^{P-one}, w_i^{P-one}, h_i^{P-one}]$), regardless of the foot type f. As the key signals for the calculation of spatiotemporal gait parameters (e.g., LHS and RHS points), these can be observed from the x^{P-one} and y^{P-one} trajectories.

To achieve feet localization, we propose a two-stage *FootChaser* framework comprised of two ConvNets: 1) *FootRegionProposer* and 2) *LocomoNet*. The FootRegionProposer proposes $n \in \mathbb{N}$ bounding boxes as 'proposed foot regions', or $PFR_{j,i}$, $j = \{1,...,n\}$ in the i^{th} frame. As there may be several false positives in the proposed regions, we hypothesized that the FootRegionProposer results may be boosted by applying another ConvNet, called LocomoNet, trained to be sensitive to the periodic/specific movement patterns embedded in the user's feet regions during gait. In other words, the LocomoNet is expected to filter out false positives by selecting the most confident regions. After applying the LocomoNet on $PFR_{j,i}$, only the frames with a single PFR are used for step width estimation (see Fig. 2).

2.1 FootRegionProposer

The FootRegionProposer is a ConvNet fine-tuned to propose PFRs in a frame. The j^{th} proposed region is in the form of a bounding box $PFR_{j,i} = [x_{j,i}, y_{j,i}, w_{j,i}, h_{j,i}]$, where $x_{j,i}$, $y_{j,i}$, $w_{j,i}$, and $h_{j,i}$ denote the center coordinates, and width and height of the box, respectively (see sample PFRs marked by red rectangles in Fig. 3). The training procedure for the LocomoNet is discussed in subsection 3.2. As noted above, there are several factors that may challenge the performance of the FootRegionProposer: 1) occlusion or extreme illumination conditions can increase the number of false negatives, 2) objects or terrain similar to the feet (i.e., noise, see Fig. 4-c), and 3) motion blur from fast movements. In addition to incorporating a fast and precise object localization/detection ConvNet (e.g., faster R-CNN [43], or YOLO [8]), a second ConvNet was applied to the FootRegionProposer output to filter false PFRs (subsection 2.2).

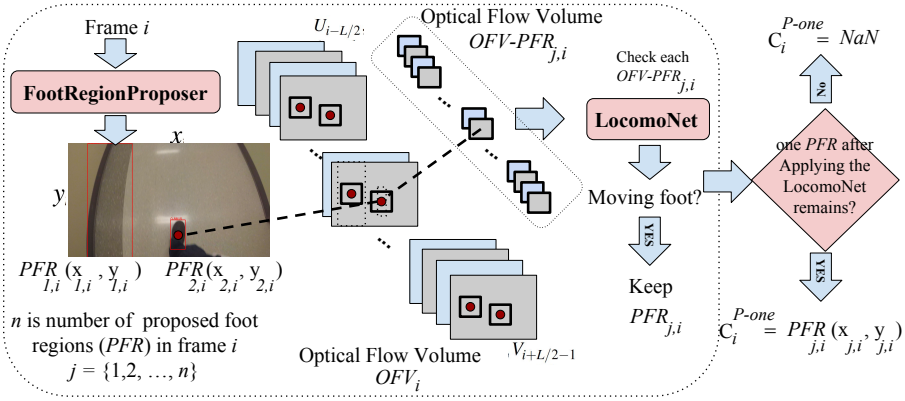


Fig. 3. The FootChaser framework. First, the FootRegionProposer proposes $n \in \mathbb{N}$ $PFR_{j,i}$ bounding boxes (red boxes), $j = \{1,2,...,n\}$) in the i^{th} frame. Multiple regions proposed are examined by LocomoNet to filter out false positives. After obtaining the stacks of optical flow volume OFV_i (V and U are vertical and horizontal 2D flow components) from the $[i - L/2, i + L/2 - 1]$ frames (L denotes the depth/length of stack), LocomoNet inputs are obtained by cropping fixed size regions centered at the center of each $PFR_{j,i}$, i.e., $(x_{j,i}, y_{j,i})$, which creates the optical flow volumes from PFRs ($OFV-PFR_{j.i}$). Final FootChaser outputs reflect frames with a single proposed region ((C_i^{P-one})).

2.2 LocomoNet: Learning from gait patterns

To reduce the number of proposed false positives (i.e., false PFRs) by FootRegionProposer Network (towards the goal of 'one' true PFR), the dynamic temporal structure of the $PFR_{j,i}$ will be further examined by the proposed LocomoNet

ConvNet. Inspired by Simonyan and Zisserman's work [5], we consider examining optical flow features to deliver bounding boxes with higher confidence of representing feet.

The horizontal $U = \{U_1, U_2, ..., U_{N-1}\}$ and vertical optical flow $V = \{V_1, V_2, ..., V_{N-1}\}$ can be calculated separately for each two consecutive frames in the video sequence (the height and width of the U and V components are equal to the frame's 2D dimension, i.e., 1080×1920). Considering a fixed length of L consecutive frames, the optical flow volume $OFV_i = \{U_{i-L/2}, V_{i-L/2}, ..., U_{i+L/2-1}, V_{i+L/2-1}\}$ is obtained for the i^{th} frame. In order to represent the temporal information of $PFR_{j,i}$, a fixed $(W_c \times H_c)$ region centered at $(x_{j,i}, y_{j,i})$ is cropped from OFV_i, which ends up to a $(2L \times W_c \times H_c)$ volume of interest $(OFV - PFR_{j,i})$ corresponding to that proposal (see Fig. 3). Each of these volumes are fed into the LocomoNet for filtering. The training procedure for LocomoNet is discussed in subsection 3.3. After applying the LocomoNet, if the output frame has only one remaining FPR, the center of that $PFR_{j,i}$ will be saved in the center vector (C_i^{P-One}). Otherwise, the corresponding component will be replaced by NaN and will not be considered in the evaluation.

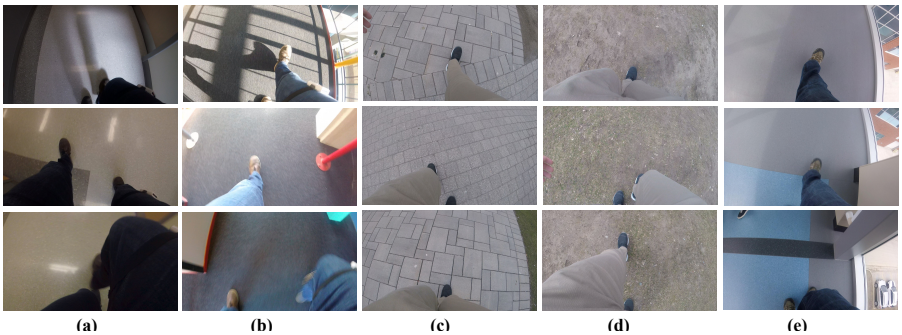


Fig. 4. Sample frames reflecting high inter-class and intra-class variability in terms of: 1) intense illuminations conditions and shadows (row 1-a,b), 2) different phases of gait, 3) different walking surfaces, e.g., color, texture (each column corresponds to a specific environment and walking surface), and 4) motion blur during crossover and side-step compensatory reactions (row 3-a,b).

3 Experiments

3.1 Dataset

Sufficiently large datasets are challenging to collect, often the primary bottleneck for deep learning. As there are no publicly available datasets specific to our needs, we employed large open datasets to initially train the FootRegionProposer and collected novel data for further training and evaluation.

For the FootRegionProposer, there are no available datasets with outside-in top-down view images of the feet from different people with a considerable diversity in appearance (e.g., shoes with different colors, shape, barefoot, socks) and movement (i.e., gait). To facilitate training, we decided to fine-tune [6] the ConvNet based on real images with normal optics from large scale datasets, which also boosts the generalizability of the network. We fine-tuned the ConvNet on Footwear (footgear) subcategory images ($\approx$ 1300 images (with bounding boxes), and 446 images of shoes from top-down view (with and without bounding boxes, and we added the bounding boxes manually)) from the ImageNet 2011 [46] dataset. Such images resemble more realistic appearance of one's footwear from different views (compared to alternatives such as UT-Zap50K [17]).

In our dataset, 3 healthy young participants (researchers affiliated with the University of Waterloo) participated in our data collection procedure. The FPV data was collected, using a GoPro Hero 5 Session camera centered on participants' belt (30fps, 1080×1920), with no specific calibration and setup. A wearable IMU was attached as closely as possible to the camera to collect movement signals (for future experiments). Overall, 5 datasets (including 2 separate data from 2 participants in different environments) were captured in five different indoor (tiles, carpet) and outdoor environments (bricks, grass/muddy) around the University of Waterloo campus, resulting in 4505 ($= 5 \times 901$) total frames (Fig. 4 shows samples from the dataset). Frames were annotated by drawing bounding boxes around the right and left shoes (in PASCAL VOC format), using the LabelImg tool [48].

In addition to the normal walking sequences, two participants were asked to simulate compensatory balance reactions (CBRs: lateral sidestep, crossover, and trip-like stepping) during gait (see Fig. 4-row 3 columns a,b for sets 1 and 2, and the GT plot for dataset 2 in Fig. 6). CBRs (near falls) are reactions to recover stability following a loss of balance (see Fig. 1-panels f and g), characterized by rapid step movements (or reaching) to widen the base of support. CBRs also introduce more challenge to our dataset as the corresponding FPV data is usually blurry (i.e., fast foot displacement) (see Fig. 4) and the field of view may be occluded.

3.2 FootRegionProposer Training

There are several models that can be taken into account for FootRegionProposer weight initialization, including SSD (Single Shot MultiBox Detector) [42], faster R-CNN [43], R-FCN [44]. In [45], it is shown that SSD models typically have (very) poor performance on small objects, such as relatively small feet regions. Among related approaches, YOLO [8] shows state-of-the-art results in terms of speed and accuracy.

To implement the FootRegionProposer, the original YOLO version 2 from the Darknet deep learning framework was used [8]. The pre-trained weights on the large-scale ImageNet dataset were used for network initialization, which was then fine-tuned on ImageNet shoe sub-category. The ConvNet was further fine-tuned on images of shoes that are captured in realistic scenes from a top-down

view. This experiment aims to advance higher detection accuracy, as they more resemble the foot regions in our FPV data. The FootRegionProposer used 1290 and 138 images for training and testing, respectively. All of the network inputs were resized to $K \times 3 \times 832 \times 832$, where $K = 64$ was the batch size (mini-batch size: 32). Moreover, the stochastic gradient descent with momentum was used as optimization method, with an initial learning rate of $\gamma = 0.001$, momentum: 0.9, and decay rate of 0.0005 (at steps 100 and 25000) selected using a Nvidia Titan X GPU. To further address the problem of limited data, the data was augmented (i.e., random crops and rotation) to improve the generalization of the network.

3.3 LocomoNet training

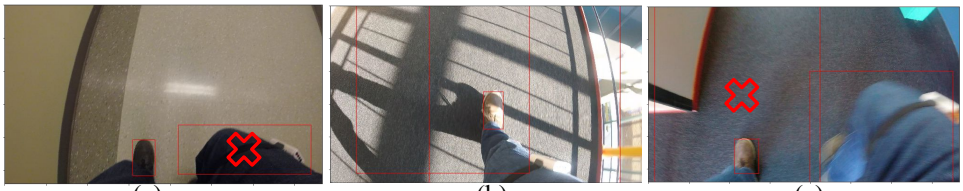
Although YOLO is very fast, it often suffers from a high number of false positives. The goal of the LocomoNet is to improve FootChaser performance by reducing the number of false proposals. The LocomoNet output maps each OFV to one of the two possible classes. Similar to [4, 1, 32], the TVL1 optical flow algorithm [47] is chosen, with OpenCV GPU implementation. Moreover, similar to [5, 32, 1], the stack length of $L = 10$ (i.e., 20 input modality channels for LocomoNet) is selected, and crop size is set to $W_c = H_c = 224$.

Based on our experiments, a 224×224 region and the stack length of $L = 10$ provided sufficient temporal information for foot regions during gait. Moreover, we handled off-the-frame crops by shifting the 224×224 box in the opposite direction in place of resizing to retain the aspect ratio. To train the LocomoNet, 300 positive (shoe/foot regions) volumes were extracted for left and right feet in each dataset in the GT data, for a total of 3000=$2 \times 300 \times 5$ true positive regions. An equal number of negatives (3000) were also randomly cropped from the non-shoe regions from the frames, with a constraint of $IoU \approx 0$ with the shoe regions at the i^{th} frame, the past and next frames in the volume were not constrained to allow for a more realistic evaluation.

The approach proposed in [32], where the authors demonstrated the possibility of pre-training temporal nets with ImageNet model, was applied in the current study. After extracting optical flow fields and discretizing the fields into [0, 255], the authors averaged the ImageNet model filters of first layer across the channel to account for the difference in input channel number for temporal and spatial nets (20 vs. 3), then copied the average results 20 times as the initialization of temporal nets. Considering such an approach, a motion stream ConvNet (ResNet-101 [50] architecture) pre-trained on video information in UCF101 dataset was used, with stochastic gradient descent and cross entropy loss. Batch size, initial learning rate, and momentum were set to $K = 64$, 0.01, and 0.9, respectively.

3.4 Evaluation

1) Model generalizability. To evaluate the extent to which *subject-related movement patterns* in *different environments* can be handled by LocomoNet, a



(a) (b) (c)

Fig. 5. Example FootRegionProposer results (PFRs) for three frames marked by red boxes. Correct foot regions were identified by the FootRegionProposer; however, false positives were also proposed. After applying the LocomoNet, some false positives were filtered out (marked with (×)). In (a) and (c) false positive(s) are successfully removed, (b) shows a case of intense illumination and shadows challenging LocomoNet, resulting two false positives that were not filtered out.

leave-one-dataset-out (LODO) cross-validation was performed. To achieve this, a $LocomoNet_{N_D}$ ($N_D = \{1, 2, ...5\}$) model was trained using the whole dataset except N_D dataset (i.e., 4800 volumes for training) and tested on the dataset N_D (i.e., 1200 volumes for testing), and repeated 5 times. The following LODO accuracies were obtained for our 5 datasets: 92.41%, 91.16%, 98.33%, 83.83%, 96.25%. The high accuracies indicate the generalizability of LocomoNet to discriminate foot-related $OFV - PFR$ in unseen datasets. The following average IoU scores were obtained for each set: 1: 0.7626, 2: 0.7304, 3: 0.3794, 4: 0.7155, and 5: 0.5235. Considering an IoU threshold of 0.5 is typically used in object detection evaluation to determine whether detection is positive (IoU of true positive> 0.5) [51], we interpret that the generalizability of the model except for $N_D = 3$, is satisfactory. We attributed the lower performance of the network on dataset 3 to the patterns of walking surface (tiles with different sizes, see Fig. 4-c).

2) The number of proposed regions with $IoU < 0.2$ (false positives) dramatically reduced after applying the LocomoNet on FPRs. To assess the false positive removal performance of the $LocomoNet_{N_D}$, we define a elimination rate metric as $ER_{N_D} = \frac{\text{Number of filtered PFRs in a specific IoU interval}}{\text{Total number of PFRs in a specific IoU interval}} \times 100$, ($IoU$=$Area(GT \cup P)/Area(GT \cap P)$). As shown in Table 1, the PFRs in a low IoU score range ($\in [0, 0.2)$), representing false positives, were removed with a high rate (e.g., in $IoU_{[0,0.1)}$ with 83.25% reduction). The relatively low true positive removal score (i.e., in $IoU_{[0.9,1)}$ with 8.09% reduction) reflects satisfactory performance of LocomoNet in retaining the true positives (refer to Fig. 2 for some failure and success cases).

3) FootChaser prediction trajectories closely match ground truth trajectories. The performance of the FootRegionProposer in tracing the GT data can be assessed by measuring 1) The individual IoU scores, and 2) the pixel-wise distance (error, E) between the the predicted foot center and its corresponding point in GT data.

As discussed in section 2, the performance of the FootChaser framework can be assessed by comparing the predicted $P - One$ bounding boxes with the $GT - one$ ($E(a^{P-One}, a^{GT-One}), a = \{x, y\}$), where mean absolute error (MAE)

Table 1. Number of proposed foot regions (N_{PFR,N_D}) and elimination rate (ER) in different intersection-over-union (IoU) intervals indicating LocomoNet ability to remove false positives by dataset. N_{PFR,N_D} dramatically reduced after applying the LocomoNet. ER_T is the weighted average of elimination rate, $IoU > 0.5$ and < 0.5, representing the true and false positives, respectively [51].)

	IoU									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$N_{PFR,1}$	1219	36	7	4	11	22	114	218	312	110
$N_{PFR,2}$	654	10	2	3	10	26	122	282	277	76
$N_{PFR,3}$	781	0	4	12	13	35	89	156	116	15
$N_{PFR,4}$	1225	2	2	1	6	31	119	293	294	36
$N_{PFR,5}$	229	18	17	27	55	106	188	195	83	10
$N_{PFR,T-}$	4108	66	32	47	95	220	632	1144	1082	247
ER1	73.83	55.55	42.85	0.00	0.00	4.54	4.38	8.25	7.05	1.81
ER2	92.20	100.00	0.00	0.00	10	11.53	13.11	17.37	13.35	10.52
ER3	97.18	100.00	0.00	8.33	7.69	5.71	0.00	1.28	3.44	6.66
ER4	83.91	50.00	100	100.00	16.66	35.48	31.93	27.30	26.87	19.44
ER5	83.40	77.77	0.00	0.00	0.00	0.00	3.72	4.61	8.43	20.00
ER_T	83.25	68.18	15.62	2.14	3.15	7.72	9.82	13.81	13.77	8.09

is taken into account as the error metric E. (see Table 2). For $GT-Two$ (e.g.,

Table 2. Mean absolute error (MAE) results for the $GT-One$ region in absolute pixels and as a fraction of image resolution. MAE $= 1/N \sum |GT - One_{a,f,i} - P - One_{a,i}|$, where $a = \{x, y\}$, $f = \{left, right\}$, $N = length(GT - One)$. MAE/R as a fraction of image resolution, where (R): R_x=1920, R_y = 1080.

	MAE (pixel)				MAE/R			
Dataset	x_{Left}	x_{Right}	y_{Left}	y_{Right}	x_{Left}	x_{Right}	y_{Left}	y_{Right}
D_1	41.68	87.50	55.66	54.81	0.021	0.045	0.051	0.050
D_2	32.90	44.00	54.29	55.94	0.017	0.022	0.050	0.051
D_3	125.74	194.85	75.19	154.46	0.065	0.101	0.069	0.143
D_4	64.40	62.57	76.11	74.11	0.059	0.070	0.057	0.068
D_5	99.31	37.68	101.52	92.04	0.051	0.019	0.094	0.085

the black dotted parts in Fig. 2), the performance was evaluated by comparing the a_i^{P-One} with the nearest GT point regardless of the foot type (Table 3 displays the results). At first glance, this may appear to be a weak metric. As discussed in section 2 and depicted in Fig. 6 and 2, in $GT - Two$ data, the FootChaser is biased toward proposing regions corresponding to the nearly-full-view feet (rather than partially-observable ones). In this application, the

Table 3. Mean absolute error (MAE) for $GT - Two$ regions in absolute pixels and as a fraction of resolution (MAE/R), where (R:) R_x=1920, R_y = 1080..

	MAE (pixel)		MAE/R	
Dataset	x	y	x	y
D_1	58.11	84.00	0.030	0.077
D_2	36.12	80.44	0.018	0.074
D_3	121.47	117.78	0.063	0.109
D_4	103.55	94.90	0.053	0.087
D_5	25.28	101.52	0.013	0.094

observed bias to larger objects is a strength as it predicts the center of the foot required for the extraction of spatiotemporal gait parameters. This can be attributed to the fact that the $FootRegionProposer$ is trained on ImageNet dataset that mainly includes the full-view images of feet. Moreover, this is in line with the findings of [20, 45], where a higher performance was reported for the detection of bigger objects in videos. Considering these points, the error criteria for $GT - Two$ regions seem to be a satisfactory representation of performance.

In addition to the relatively low error rates ($< 10\%$ for the x trajectories), as presented in Fig. 6, the framework also predicted many of the points at the timings of CBRs (spikes). Therefore, these trajectories can be a promising avenue for the detection of CBRs. High E values for D_3 (Tables 2 and 3) also support the low IoU rate achieved for that dataset (due to the patterns of the walking surface).

4 Conclusion and future work

As the main contribution, this study addressed the potential of incorporating a body-mounted camera to develop automated markerless algorithms to detect feet in natural environments. This advances our long-term objective to develop novel markerless models to extract spatiotemporal gait parameters, particularly step width, to complement existing IMU-based methods.

As the next steps, we aim to: 1) collect synchronized criterion (gold) standard human movement data using motion capture (e.g., Vicon) or gait analysis tools (e.g., pressure-sensitive mat, GaitRite) synchronized to FPV data and develop a model to convert the pixel-wise results of the FootChaser into the commonly-used distance units (e.g., m or cm), and 2) develop a more robust version of FootChaser framework by collecting a larger free-living dataset from older adults with different frailty levels, annotate them, and make the data publicly available.

This paper contributes an advance in the field of ambulatory gait assessment to localize feet in a waist-mounted FPV feed towards a fully automatic system to detect abnormalities (e.g., compensatory balance reactions, or near-falls), identify environmental hazards (e.g., slope changes, stairs, curbs, ramps)

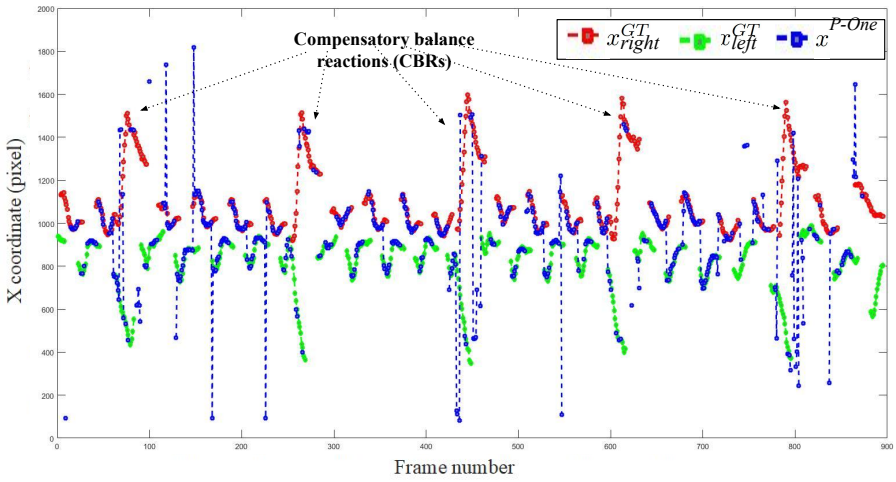


Fig. 6. Time series plot of X coordinate center of the most confident proposed foot regions (PFR, blue) predicted by the FootChaser framework for dataset 2. Ground truth (GT) for the left and right feet are plotted in green and red, respectively. Spikes represent compensatory balance reactions (CBRs) performed by the participant.

and surfaces (e.g., gravel, grass, concrete) that influence mobility and potential risk to falls. As described earlier, FPV data also provides objective evidence on cause and circumstances of perturbed balance during activities of daily living, Our future studies will examine the potential for automatic detection of these environmental fall risk hazards [12, 11].

Given massive amounts of unlabeled FPV data collected during longer-term study, we aim to develop approaches that can robustly handle significant diversity in movement patterns (e.g., rhythm, speed), different populations (e.g., older fallers, Alzheimer's disease), and varying clothing and footwear appearance. To address these aspects, we aim to *personalize* both of the FootRegionProposer and LocomoNet ConvNets to introduce an adaptive pipeline *"AdaFootChaser"* similar to [39] in our future work.

Acknowledgments. Research supported by National Sciences and Engineering Research Council of Canada (NSERC), and by AGE-WELL NCE Inc. M. Nouredanesh was funded by an AGE-WELL Inc. (Canadas technology and aging network) Graduate Scholarship.

References

1. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1894-1903

2. Abebe, G., & Cavallaro, A. (2017, October). A long short-term memory convolutional neural network for first-person vision activity recognition. In Proc. of International Conference on Computer Vision Workshops (ICCVW), Venice, Italy (Vol. 1, No. 2, p. 3).
3. Abebe, G., & Cavallaro, A. (2017, October). Inertial-Vision: cross-domain knowledge transfer for wearable sensors. In Proc. of International Conference on Computer Vision Workshops (ICCVW), Venice, Italy (Vol. 7).
4. Song, S., Chandrasekhar, V., Mandal, B., Li, L., Lim, J. H., Sateesh Babu, G., ... & Cheung, N. M. (2016). Multimodal multi-stream deep learning for egocentric activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 24-31).
5. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).
6. Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1717-1724).
7. Lord, S., Galna, B., Verghese, J., Coleman, S., Burn, D., & Rochester, L. (2012). Independent domains of gait in older adults and associated motor and nonmotor attributes: validation of a factor analysis approach. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 68(7), 820-827.
8. Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. arXiv preprint, 2017.
9. Taylor, K., Reginatto, B., Patterson, M.R., Power, D., Komaba, Y., Maeda, K., Inomata, A. and Caulfield, B., 2015, August. Context focused older adult mobility and gait assessment. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE (pp. 6943-6946). IEEE.
10. Hickey, A., Del Din, S., Rochester, L., & Godfrey, A. (2016). Detecting free-living steps and walking bouts: validating an algorithm for macro gait analysis. Physiological measurement, 38(1), N1.
11. Nouredanesh, M., McCormick, A., Kukreja, S. L., & Tung, J. (2016). Wearable Vision Detection of Environmental Fall Risks using Convolutional Neural Networks. arXiv preprint arXiv:1611.00684.
12. Nouredanesh, M., McCormick, A., Kukreja, S. L., & Tung, J. (2016, June). Wearable vision detection of environmental fall risk using Gabor Barcodes. In Biomedical Robotics and Biomechatronics (BioRob), 2016 6th IEEE International Conference on (pp. 956-956). IEEE.
13. Iluz, T., Gazit, E., Herman, T., Sprecher, E., Brozgol, M., Giladi, N., ... & Hausdorff, J. M. (2014). Automated detection of missteps during community ambulation in patients with Parkinson's disease: A new approach for quantifying fall risk in the community setting. Journal of neuroengineering and rehabilitation, 11(1), 48.
14. Brodie, M. A., Lord, S. R., Coppens, M. J., Annegarn, J., & Delbaere, K. (2015). Eight-week remote monitoring using a freely worn device reveals unstable gait patterns in older fallers. IEEE Transactions on Biomedical Engineering, 62(11), 2588-2594.
15. Jiang, H., & Grauman, K. (2017, July). Seeing invisible poses: Estimating 3D body pose from egocentric video. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 3501-3509). IEEE.
16. Huang, Yinghao, Federica Bogo, Christoph Classner, Angjoo Kanazawa, Peter V. Gehler, Ijaz Akhter, and Michael J. Black. "Towards Accurate Markerless Human Shape and Pose Estimation over Time." arXiv preprint arXiv:1707.07548 (2017).

17. A. Yu and K. Grauman. "Fine-Grained Visual Comparisons with Local Learning". In CVPR, 2014.
18. Brattoli, B., Buchler, U., Wahl, A. S., Schwab, M. E., & Ommer, B. (2017, January). Lstm self-supervision for detailed behavior analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 1).
19. Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., ... & Theobalt, C. (2017). MARCOnI-ConvNet-Based MARker-less motion capture in outdoor and indoor scenes. IEEE transactions on pattern analysis and machine intelligence, 39(3), 501-514.
20. A. Rozantsev, V. Lepetit, and P. Fua. Flying Objects Detection from a Single Moving Camera. In CVPR, 2015.
21. Jain, A., Tompson, J., LeCun, Y., & Bregler, C. (2014, November). Modeep: A deep learning framework using motion features for human pose estimation. In Asian conference on computer vision (pp. 302-315). Springer, Cham.
22. Tekin, B., Rozantsev, A., Lepetit, V., & Fua, P. (2016). Direct prediction of 3d body poses from motion compensated sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 991-1000).
23. Rogez, Grrégory, James S. Supančič, and Deva Ramanan. "First-person pose recognition using egocentric workspaces." In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pp. 4325-4333. IEEE, 2015.
24. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., & Theobalt, C. (2017, April). Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of International Conference on Computer Vision (ICCV) (Vol. 10).
25. Hu, Richard Zhi-Ling, et al. "3D Pose tracking of walker users' lower limb with a structured-light camera on a moving platform." Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. IEEE, 2011.
26. Ng, S., Fakih, A., Fourney, A., Poupart, P., & Zelek, J. (2009, September). Towards a mobility diagnostic tool: Tracking rollator users' leg pose with a monocular vision system. In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE (pp. 1220-1225). IEEE.
27. Page, S., Martins, M. M., Saint-Bauzel, L., Santos, C. P., & Pasqui, V. (2015, May). Fast embedded feet pose estimation based on a depth camera for smart walker. In Robotics and Automation (ICRA), 2015 IEEE International Conference on (pp. 4224-4229). IEEE.
28. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H. P., & Theobalt, C. (2018). Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. arXiv preprint arXiv:1803.05959.
29. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H. P., ... & Theobalt, C. (2016). Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics (TOG), 35(6), 162.
30. von Marcard, T., Rosenhahn, B., Black, M. J., & PonsMoll, G. (2017, May). Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In Computer Graphics Forum (Vol. 36, No. 2, pp. 349-360).
31. Shiratori, T., Park, H.S., Sigal, L., Sheikh, Y., Hodgins, J.K.: Motion capture from body-mounted cameras. ACM Transactions on Graphics 30(4) (2011) 31:110
32. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L., 2016, October. Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision (pp. 20-36). Springer, Cham.

33. Phillips, L. J., DeRoche, C. B., Rantz, M., Alexander, G. L., Skubic, M., Despins, L., ... & Koopman, R. J. (2017). Using embedded sensors in independent living to predict gait changes and falls. Western journal of nursing research, 39(1), 78-94.
34. Bigy, A. A. M., Banitsas, K., Badii, A., & Cosmas, J. (2015, April). Recognition of postures and Freezing of Gait in Parkinson's disease patients using Microsoft Kinect sensor. In Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on (pp. 731-734). IEEE.
35. Babak Taati PhD, P., & Alex Mihailidis PhD, P. (2014). Vision-based approach for long-term mobility monitoring: Single case study following total hip replacement. Journal of rehabilitation research and development, 51(7), 1165.
36. Gabel, M., Gilad-Bachrach, R., Renshaw, E., & Schuster, A. (2012, August). Full body gait analysis with Kinect. In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE (pp. 1964-1967). IEEE.
37. Cippitelli, E., Gasparrini, S., Spinsante, S., & Gambi, E. (2015). Kinect as a tool for gait analysis: validation of a real-time joint extraction algorithm working in side view. Sensors, 15(1), 1417-1434.
38. Auvinet, E., Multon, F., Manning, V., Meunier, J., & Cobb, J. P. (2017). Validity and sensitivity of the Longitudinal Asymmetry Index to detect gait asymmetry using Microsoft Kinect data. Gait & posture, 51, 162-168.
39. Charles, J., Pfister, T., Magee, D., Hogg, D., & Zisserman, A. (2016, June). Personalizing human video pose estimation. In Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (pp. 3063-3072). IEEE.
40. Guler, Riza Alp, Natalia Neverova, and Iasonas Kokkinos. "Densepose: Dense human pose estimation in the wild." arXiv preprint arXiv:1802.00434 (2018).
41. Wang, Y., Liu, Y., Tong, X., Dai, Q., & Tan, P. (2018). Outdoor markerless motion capture with sparse handheld video cameras. IEEE transactions on visualization and computer graphics, 24(5), 1856-1866.
42. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.
43. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
44. J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. arXiv preprint arXiv:1605.06409, 2016. 1, 2, 3, 4, 5, 6
45. Huang, Jonathan, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer et al. "Speed/accuracy trade-offs for modern convolutional object detectors." In IEEE CVPR. 2017.
46. Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.
47. Zach, Christopher, Thomas Pock, and Horst Bischof. "A duality based approach for realtime TV-L 1 optical flow." Joint Pattern Recognition Symposium. Springer, Berlin, Heidelberg, 2007.
48. Tzutalin. LabelImg. Git code (2015). https://github.com/tzutalin/labelImg
49. Brach, Jennifer S., et al. "Too much or too little step width variability is associated with a fall history in older persons who walk at or near normal gait speed." Journal of neuroengineering and rehabilitation 2.1 (2005): 21.
50. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

51. Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303-338.
52. Ihlen, E. A., Weiss, A., Bourke, A., Helbostad, J. L., & Hausdorff, J. M. (2016). The complexity of daily life walking in older adult community-dwelling fallers and non-fallers. Journal of biomechanics, 49(9), 1420-1428.