Automated recognition of individual performers from de-identified video sequences

Machine Learning with Applications 11 (2023) 100450

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Automated recognition of individual performers from de-identified video sequences

Zizui Chen [a], Stephen Czarnuch [a,b,*], Erica Dove [c], Arlene Astell [c,d,e]

[a] Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, 320 Elizabeth Ave., St. John's, A1C 4Z6, Newfoundland, Canada
[b] Discipline of Emergency Medicine, Memorial University of Newfoundland, 320 Elizabeth Ave., St. John's, A1C 4Z6, Newfoundland, Canada
[c] KITE – Toronto Rehabilitation Institute, University Health Network, 550 University Avenue, Toronto, ON, M5G 2A2, Canada
[d] Department of Occupational Sciences & Occupational Therapy and Department of Psychiatry, University of Toronto, Toronto, M5G 1V7, Canada
[e] School of Psychology & Clinical Language Science, University of Reading, Reading, RG6 7BE, UK

ARTICLE INFO

Keywords:
Human identification
Differentiation
Discrimination
Depth image processing
Skeleton data
Privacy

ABSTRACT

Identification of individual humans from RGB image data is well-established. However, in many domains, such as in healthcare or applications involving children, ethical issues have been raised around using traditional RGB image data because individuals can be identified from these data. The widespread availability of reliable depth data, and the associated human skeleton data derived from these data, presents an opportunity to differentiate between individuals while potentially avoiding individually identifiable features.

Using skeleton data only, we developed a unique 20-dimensional bone segment length feature vector for 1,761 trials (1,759,980 image frames) of data, captured from 14 participants who engaged in a one-hour group intervention playing Xbox One Kinect Bowling twice-weekly for 24 weeks. We then evaluated our novel feature using representative batch processing (k-nearest neighbour) and real-time (multi-layer perceptron) models, validated against manually-labelled ground-truth data. Our results suggest that our skeleton feature can differentiate between instances (i.e., individuals) with an accuracy over all participants of 100% for batch processing and 96.57% in real-time, and deals well with class imbalances. Our results suggest that we can reliably differentiate between individual persons using only skeleton data derived from depth image data in medical research.

1. Introduction

Human recognition algorithms commonly use cameras to capture colour images and identify features, such as eigen-face (Gonzalez & Woods, 2006). Recently, deep learning approaches have been widely used in related areas, such as face recognition (Khan, Harous, Hassan, Ghani Khan, Iqbal, & Mumtaz, 2019; Rathgeb, Dantcheva, & Busch, 2019) and action recognition (Wang, Xu, Cheng, Xia, Yin, & Wu, 2018). In medical contexts, these recognition technologies (Ding & Tao, 2015; Sun, Liang, Wang, & Tang, 2015) often raise privacy concerns, because the collected images are stored and processed by multiple computer systems and multiple operators. All data necessarily include patients' identifiable features, such as faces, and it is difficult to prevent unauthorised access or duplication. To make things worse, these digital images are often connected to patients' physical health and performance data (Parajuli, Tran, Ma, & Sharma, 2012; Saha, Pal, Konar, & Janarthanan, 2013). Accordingly, medical evaluations are traditionally performed one-on-one with physicians and patients, which is inefficient and expensive. Still, despite existing challenges, the potential benefits of automated recognition of individuals is clear, motivating significant research into computer vision technologies that can semi- or fully-automatically identify individuals and perform various forms of automated evaluation (Hoey et al., 2010; Ng et al., 2020), intended to enhance clinical outcomes. Most of these existing technologies use cameras and wearable sensors (Teipel et al., 2018), but to our knowledge, these technologies can only be used to evaluate patient clinical performance after explicitly identifying the patient.

In this paper, we address the human recognition problem in clinical applications by proposing a new framework to identify and differentiate between humans with privacy in mind. Our approach temporarily processes paired colour and depth images of patients (e.g., from a commercially available depth camera like the Microsoft Kinect (Microsoft Kinect Developer, 2017)), and from the images we generate a special feature that can differentiate and discriminate between individuals — the 3D skeleton. Skeleton data does not include any traditional identifiable information. Furthermore, colour and depth images used to generate the skeleton data are immediately discarded after the

[*] Corresponding author.
E-mail address: sczarnuch@mun.ca (S. Czarnuch).

https://doi.org/10.1016/j.mlwa.2023.100450
Received 9 November 2022; Received in revised form 17 January 2023; Accepted 17 January 2023
Available online 21 January 2023

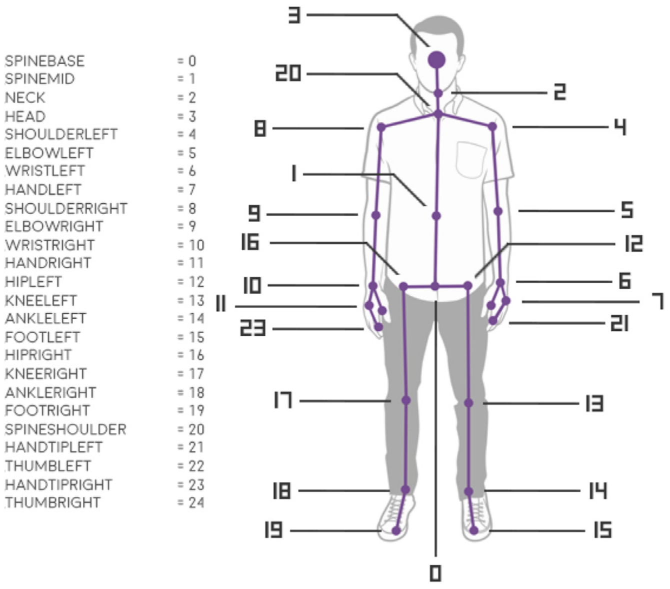

Fig. 1. Microsoft Kinect software development kit skeleton data features (Sapinski, Kaminska, Pelikant, & Anbarjafari, 2019). Original source (https://msdn.microsoft.com) no longer available.

skeleton is generated, allowing us to preserve differentiating knowledge without retaining identifiable information. The drawback of using these anonymised skeleton data is that traditional human recognition algorithms are no longer applicable. Thus, we respond to the challenge of recognising individual humans and differentiating between humans from this novel and non-traditional skeleton feature. We first design our new skeleton feature based on 3D bone-lengths. Then we use statistics and machine learning-based approaches to show the discriminative nature of our novel feature for humans recognition.

2. Literature review

Human recognition and segmentation tasks such semantic labelling of humans (Czarnuch & Mihailidis, 2016; Garcia-Garcia, Orts-Escolano, Oprea, Villena-Martinez, & Garcia-Rodriguez, 2017; Zhao, Feng, Wu, & Yan, 2017), eye tracking (Krafka et al., 2016), pose estimation (Shotton et al., 2013; Xiang, Schmidt, Narayanan, & Fox, 2017), facial recognition (Jiang & Learned-Miller, 2017), human segmentation (Liang et al., 2016; Zhang et al., 2019), gesture recognition (Kim, Lee, & Park, 2008), object detection for autonomous driving (Feng et al., 2019), and human tracking (Chu et al., 2017; Fan, Xu, Wu, & Gong, 2010) have been widely researched with good results. However, these technologies require direct access to colour (i.e., RGB) images, which pose privacy concerns in medical research environments.

The widespread availability of reliable depth and RGB (RGB-D) data from sensors such as the Microsoft Kinect (Microsoft Kinect Developer, 2017) has seen burgeoning development of 3D vision-based analysis in research, such as gait analysis (e.g., Preis, Kessel, Werner, & Linnhoff-Popien, 2012), posture recognition, pose estimation (e.g., Le, Nguyen, et al., 2013), and ground plane detection (e.g., Zhang & Czarnuch, 2020). Arguably, one of the most significant advantage of RGB-D data over RGB data is the ability to reliably generate full 3D skeleton representations of humans visible in the scene. These skeleton data (see Fig. 1) allow real-time human segmentation and subsequent 3D analysis of human motion, with most applications focused on pose estimation and action recognition (Cippitelli, Gasparrini, Gambi, & Spinsante, 2016; Czarnuch & Mihailidis, 2016; Li, Hou, Wang, & Li, 2017; Liu, Akhtar and Mian, 2017; Liu, Wang, Duan, Abdiyeva, & Kot, 2018).

The concurrent availability of depth, skeleton, and RGB data have allowed the development of new techniques that combine the strengths of existing 2D image processing techniques with the novelty and power of more recent advances in 3D movement analysis (Barmpoutis, 2013; Zhao, Liu, Yang, & Cheng, 2012). Yet, the use of these data in some domains, such as healthcare, raise ethical issues, including potential privacy concerns arising from the ability to identify individuals from RGB images (Ding & Tao, 2015; Sun et al., 2015) and connect physical health and performance data to those persons (Parajuli et al., 2012; Saha et al., 2013). Accordingly, reliably discriminating between different individuals using only data like depth images and skeleton representations (i.e., without RGB image data) would preserve the ability to perform tasks such as movement analysis or activity detection while significantly reducing the likelihood of identifying who an individual is. To our knowledge, reliably discriminating between individuals using only skeleton or depth data has not yet been attempted or accomplished.

Our objective is to utilise only anonymised skeleton data collected from RGB-D sensors located in indoor public spaces to differentiate between unique persons engaging in individual or group activities and label these individuals reliably, even over multiple sessions. In this way, an individual's movement and activities can be identified and analysed, and their performance can be assessed over consecutive visits to the spaces, without associating their identity to their label.

Various reliable techniques of identifying individuals, including contact-based approaches such as fingerprint readers and other biometrics (Jain, Ross, & Prabhakar, 2004; Maltoni, Maio, Jain, & Prabhakar, 2009), achieving identification accuracies that approach 100%. Non-contact techniques, such as retinal scanning and those using images or video data, are attractive for many applications though, because people can be unobtrusively identified, ideally in real-time, without any intentional interaction, though identification accuracy is typically reduced compared to contact-based approaches. Image processing techniques, such as those using deep learning, achieve high accuracy in identifying humans from RGB images (Nakajima, Pontil, Heisele, & Poggio, 2003) when ideal conditions exist (e.g., full human faces are visible and in focus). For example, Sun et al. (2015) implemented facial recognition on the LFW face database (Huang, Mattar, Berg, & Learned-Miller, 2012), achieving an overall classification accuracy of 99.53% (Sun et al., 2015). Similarly, Lawrence et al. used a shallow CNN and achieved a 3.8% error rate (Lawrence, Giles, Tsoi, & Back, 1997).

Identifying individual humans with partial data or under non-ideal conditions represented in many practical real-world scenarios (e.g. blurred images, sub-optimal camera perspective) remains challenging for many applications, such as real-time security surveillance systems (Koo, Cho, Baek, Kim, & Park, 2018) and risk situations (Wieczorek, Siłka, Woźniak, Garg, & Hassan, 2022). Under these more challenging conditions, approaches such as using ResNet to recognise faces (Lu, Jiang, & Kot, 2018) have achieved classification accuracies of at least 92% using SCface (Garcia-Garcia et al., 2017) and LFW (Huang et al., 2012) datasets in reduced-resolution images. These results suggest that approaches of identifying humans from video data in more challenging conditions have improved significantly in recent years, particularly those using machine learning and deep learning (Vizilter, Gorbatsevich, Vorotnikov, & Kostromov, 2016; Xiang, Zhang, Tang, Zou, & Xu, 2018). However, in some circumstances, collecting video or RGB data (i.e., image sequences) is not ideal, such as in healthcare applications, situations involving children, scenarios where people may be exposed, or in conditions where informed consent is questionable (e.g., persons with cognitive impairments). Under these circumstances, it may still be useful to know who individuals are (e.g., to track disease progression over time), or at a minimum, to be able to discriminate between different people. If persons could be anonymously labelled without connecting this label to their real identity, complications (e.g., ethical and privacy issues) could be reduced while still achieving the benefits of tracking individuals over time.

Skeleton data are widely used in research. However, most research using skeleton data focus on recognising pose (e.g., Du, Wang, & Wang, 2015) and activities (e.g., Basak et al., 2022; Czarnuch & Mihailidis, 2016), rather than identifying or differentiating between individuals. Researchers use sequences of different types of skeletons to recognise human actions (e.g., skeleton by Kinect (Liu, Liu and Chen, 2017), and star skeleton (Chen, Chen, Chen, & Lee, 2006)). Machine learning is another approach to recognise human actions (Zhang et al., 2017). Other researchers build their own features out of skeletons, including joints coordinates (Gaglio, Re, & Morana, 2015; Shan & Akella, 2014), joints distances (Cippitelli et al., 2016), histogram of joints (Xia, Chen, & Aggarwal, 2012), angles between joints (Zhang & Tian, 2012) and combined skeleton with RGB data (Franco, Magnani, & Maio, 2020) to recognise human actions. Gait recognition is another application of skeletons. By using both static features (e.g., distances between selected joints) and dynamic feature (e.g., angles of swing limbs) from sequences of skeletons, researchers have used naive machine learning (k-NN) to build walking models (Sun et al., 2018). CNN-based machine learning approaches are also used in gait applications (Yao et al., 2021). Individual skeletons can be further used in forensics for recognising human age and sex (Mesejo, Martos, Ibanez, Novo, & Ortega, 2020). When using skeleton data for pose and gait recognition, both spatial and temporal information are needed, which means one single frame with a skeleton is not enough. These recognition approaches instead require a sequence of skeletons. Research on applying skeletons in human recognition is limited. For example, Sinha et al. created a handcrafted feature to recognise humans from sequences of skeletons (Sinha, Chakravarty, Bhowmick, et al., 2013). Since our goal is to recognise humans from individual frames, the above approaches cannot be applied directly. However, the handcrafted feature defined by Sinha et al. (2013) and the machine-learning based classifier by Sun et al. (2018) are inspirational to our research.

Similar to our objective, Wang, Tan, Ning, and Hu (2003) sought to identify humans using only structural changes and transitions of an individual's silhouette while walking using two different classifiers and two different similarity measures from three different camera perspectives, and report individual person classification results that varied substantially from 63.75% to 93.75%. While the results of Wang et al. (2003) were encouraging, their approach was sensitive to the camera perspective and was ineffective at recognising the same person across days. Other research on identifying humans with non-RGB data (e.g., only depth images or skeleton data) remain very limited.

The joints provided by skeleton representation of a person (e.g., Fig. 1) are typically derived from only the depth image of a depth sensor. These approaches provide the locations of anatomical body parts (e.g., head, feet) and joints (e.g., shoulders, elbows) in 3D space. We hypothesise, building on the work of Wang et al. (2003), that the set of bone segment lengths, measured as the Euclidean distance between estimated anatomical points in 3D, will remain sufficiently constant for an individual regardless of pose, and together these bone segment lengths will uniquely represent individual humans distinctly from others. Furthermore, because the skeleton representation is three-dimensional, our approach is theoretically perspective-independent, overcoming one of the most significant limitations of previous, albeit 2D, approaches. By building bone segment lengths as a vectors, it is possible to apply machine learning-based classification algorithms to uniquely model individuals and identify persons over multiple days using only their skeletons.

3. Privacy-preserved data collection

Our main objective was to set up scenarios where we could use a Microsoft Kinect RGB-D sensor to collect skeleton data from a group of diverse and unique participants performing various actions. For the purposes of our research, we also intended to collect RGB data such that we could associate the skeleton data we collected to individual participants, strictly for the purpose of evaluating the performance of our novel human recognition feature. This research is approved by our Research Ethics Board.

Table 1
Participant demographics.

Part.[a]	Sex	Age	Education	MoCA score[b]	Walking aid
00	M	80	Unsure	8	Walker
01	M	65	Grade 10	21	Cane
02	M	83	Unsure	18	No
03	F	72	Grade 12	11	No
04	M	60	Grade 12	26	No
05	M	92	Grade 12	15	Cane
06	M	80	Unsure	18	No
07	F	80	Grade 9	19	Walker
08	F	84	Grade 10	19	No
09	F	78	Grade 12	5	Wheelchair
10	F	77	Elementary	8	No
11	M	58	Grade 12	17	No
12	F	93	Grade 10	11	Walker

[a] No data for participant 13.

[b] Montreal Cognitive Assessment (MoCA) score (out of 30).

3.1. Participants and intervention

Our participants were older adults (many with dementia, mild cognitive impairment or physical impairment) who utilised the recreational activities, wellness clinics, continuing education, or other services provided by senior citizens centres in Ontario, Canada. The demographics of our participants are shown in Table 1. As part of the activities offered by one centre in Ontario, we organised 24 one-hour group bowling sessions using Kinect Sports and the Microsoft Xbox One system twice per week for 12 weeks in a large activity room. The main purpose of the intervention was to quantify whether people with dementia or MCI could learn to use the technology and improve over time (Dove & Astell, 2019). Each session was facilitated by a member of the research team, though participants were encouraged to participate in the activity as independently as possible. We configured the room so that a half-circle of chairs (of varying number depending on the number of participants during each session; diameter of 5 m) faced a large-screen TV used for the bowling intervention. We taped a line on the floor at 2 m from the TV so that the entire body of the primary participant using the Kinect was clearly and fully visible to the system. Participants generally remained seated on a chair in the half-circle, and the facilitator instructed participants to stand up and walk to the line at the centre of the scene for each trial one at a time, then return to a seat after their turn. Participants who were not currently in a trial generally remained seated, but were allowed to switch seats and engage socially with other members of the group at their discretion or leave the circle for other reasons (e.g., bathroom breaks).

We placed a second Kinect 2 sensor directly beside the gaming system sensor, and we connected the second sensor to a computer running the Microsoft Kinect SDK (Microsoft Kinect Developer, 2017). Using the SDK, we collected skeleton data from up to six participants at a time at 30 frames per second (FPS), as well as depth images at one FPS for help with video annotation. We set up two additional video recorders to capture video data that were used to assist with annotation. Only one participant was generally in the centre of the scene at any point in time. However, at times multiple people could be in the centre of the scene (e.g., the facilitator helping a participant, a participant crossing through the centre of the room to change seats). Additionally, the Kinect SDK periodically captured skeleton data from participants who were near the edges of the sensor field of view. SDK tracking data from the first session were lost because of a technical issue with our sensor.

3.2. Participant data collection

The number of trials completed by each participant at each session are shown in Table 2. Fourteen[1] participants completed a total of 1761

[1] Data from one participant were excluded from our study because of the this participant's limited participation during the period of our data collection.

Table 2
Number of trials by participant by session.

Session	Participant														Total
	00	01	02	03	04	05	06	07	08	09	10	11	12	13	
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0
2	6	6	0	6	6	6	6	6	6	6	0	0	0	6	10
3	7	7	0	7	7	7	7	7	0	7	6	7	0	0	10
4	7	7	0	0	0	7	6	7	7	6	0	0	0	0	7
5	8	0	0	8	0	8	8	8	8	0	5	8	8	0	9
6	7	7	0	7	7	7	7	7	7	7	0	0	0	0	9
7	7	7	0	6.5[a]	0	6	7	7	7	0	7	7	7	0	10
8	12	0	0	12	12	12	12	12	12	0	0	0	0	0	7
9	9	0	1	0	7	8	9	9	9	0	9	9	9	0	10
10	10	9.5[a]	0	10	10	10	10	9	9	0	0	0	0	0	8
11	9	10	0	10	10	10	10	9	10	0	0	0	0	0	8
12	10	10	0	10	10	10	10	0	10	0	10	10	0	0	9
13	14	14	0	14	14	0	14	14	0	0	0	0	0	0	6
14	8	8	0	8	8	8	8	8	8	0	8	8	8	0	11
15	12	9	0	12	0	12	12	12	12	0	0	0	0	0	7
16	11	0	0	10	0	11	11	11	11	0	11	11	0	0	8
17	13	0	0	13	13	13	13	7	13	0	0	0	0	0	7
18	10	10	0	10	10	10	10	10	10	0	10	0	0	0	9
19	11	11	0	0	11	11	11	10	11	0	0	0	0	0	7
20	8	8	0	8	8	8	8	8	8	0	8	0	4	0	10
21	12	12	0	12	12	12	12	12	0	0	0	0	0	0	7
22	10	0	0	10	0	10	10	10	0	0	0	0	0	0	5
23	9	9	0	10	9	9	7	8	0	0	8	8	8	0	10
24	13	0	0	13	13	0	13	13	0	0	12	13	0	0	7
Average trials[b]	9.7	9.0	1	9.8	9.8	9.3	9.61	9.3	9.3	6.5	8.6	9	7.3	6	9.1
Total sessions	23	16	1	20	17	21	23	22	17	4	11	9	6	1	13.6

[a] Trial was started but not completed.

[b] Average values are reported only for days when participants participated in sessions.

trials, with the average participant completing 9.1 trials per session attended ($SD = 2.40$). Participants completed at least one trial in an average of 13.6 ($SD = 8.17$) of the group sessions. The average trial was 33.1 s (992 frames) long ($SD = 10.84$ s), and we captured a total of 16 h 17 m 43 s or 1,759,890 frames of data. Since many frames had multiple tracked skeletons, we tracked a total 5,164,400 skeletons and manually labelled 1,447,559 frames where a participant was actively in the centre of the scene. We left skeletons belonging to participants other than the active participant unlabelled since they were either not properly tracked (e.g., too close to the range of the sensor) or were background participants sitting in the waiting area. We captured an average of 33 frames of depth data per trial, and a total of 58,663 frames of depth data across all sessions.

4. Human recognition with anonymised skeleton data

Traditional human recognition algorithms do not work on skeleton data. Rather, most modern approaches directly utilise RGB images. From our collected skeleton data we developed two datasets: (1) a ground-truth labelled dataset; and (2) a bone segment feature space dataset. Then, we built two separate classifiers, k-Nearest Neighbour (k-NN) and a Multi-layer Perceptron (MLP) model (see Azarloo & Farokhi, 2012; García, Mollineda, & Sánchez, 2008; Kayikcioglu & Aydemir, 2010; Pacheco & López, 2019) to demonstrate the discrimination of our new feature and to evaluate our feature's performance using representative batch processing and real-time classifiers.

4.1. Datasets

4.1.1. Ground-truth dataset

Using the data captured from the two video recorders, we manually identified the approximate start and stop times of each trial for each participant. Data from all sessions were merged to form a single dataset file. We then labelled each of the participants with a unique *participant ID* (i.e., P_n, where $n = \{0 \ldots 12\}$), and we associated this ID with the start and end time of each trial. For this manual labelling, we defined an active participant as a participant who was performing, getting ready to perform or leaving after performing the action (i.e., throwing the virtual bowling ball). We defined the start time as the timestamp of the frame when the active participant was visible in the scene and began moving toward the centre of the half-circle. We defined the stop time for each participant as the timestamp of the frame where the participant returned to their seat or left the scene, or the timestamp of the frame where the next participant's trial started. This ensured that there was only ever one active participant. We used the manually labelled times to associate the skeleton to *participant IDs*. This allowed us to connect multiple skeletons that were assigned to the same participant during a trial to a single *participant ID*.

We created a tool (Chen, 2020) to visualise the captured skeleton data and assist with the manual labelling process. A sample of the tool visualising a skeleton is in Fig. 2. With our tool, we viewed the ground-truth dataset second by second by aggregating all skeletons in one second, ensuring that each manually labelled *participant ID* was correct, and assigning the missing *participant ID* to any unlabelled skeleton by tracking the active participant forward or backward in time until the *participant ID* was found. Further, we used a majority voting method to fix mislabelled skeletons in a *tracking ID*. The resulting ground-truth dataset was comprised of 1,447,559 frames with the spatial coordinates of all 25 *Tracked* parts, a known *participant ID*, and a timestamp. Unlabelled data were removed from this dataset. A process to fix obviously mislabelled data by majority voting (i.e, skeleton with same *Tracking Id* but different *labels*) was used to correct 165,573 mislabelled data frames. Table 3 shows the number of *Tracking Ids* and tracked frames per participant.

4.1.2. Bone segment feature space dataset

Our hypothesis is that humans can be characterised, or more specifically, that humans can be differentiated by the unique combination of their bone segment lengths. From this hypothesis, we derive two assumptions: (1) that bone segment feature vectors from the same individual will remain reasonably consistent over multiple frames (i.e., bone segment lengths will always reasonably cluster); and (2) that bone segment feature vectors from different individuals will be distinct (i.e., the feature vector of bone lengths will be distinct) in feature space.

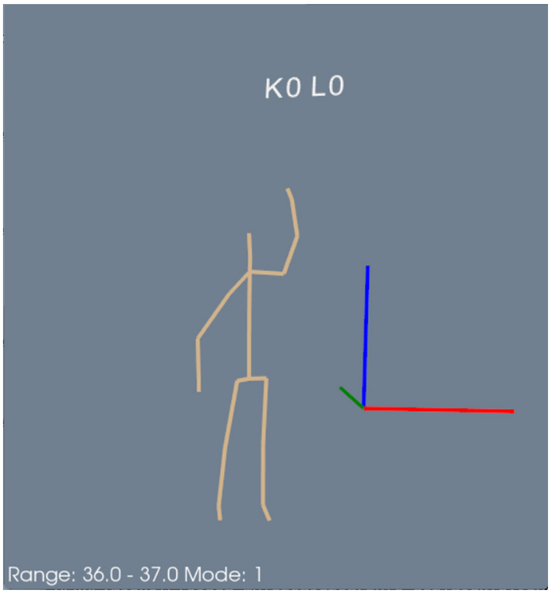

Fig. 2. Sample of a visualised skeleton.

Table 3
Tracking Ids and frames per participant.

Label	Number of tracking Id	Number of frames
0	99	115,530
1	112	196,509
2	80	85,035
3	152	159,295
4	164	153,615
5	172	230,803
6	19	43,206
7	30	39,875
8	67	84,326
9	125	190,311
10	59	48,391
11	6	8727
12	118	91,936
Total	1203	1,447,559

We further created a bone segment feature space dataset using the 3D length of each bone segment, defined by connecting anatomically related joints. Using the spatial coordinates of the 25 joints for each participant in each frame of the ground-truth dataset, we calculated the 3D lengths of 20 bone segments (Table 4). Our final bone segments and associated lengths replace the spatial coordinate data in each frame of the ground-truth dataset (Section 4.1.1), comprising our bone segment feature space dataset.

The ideal input to our classifier would be perfect skeletons. However, due to sensor error and skeleton estimation, the data are not perfect. To show the variation in our bone length estimation, we ran a statistical analysis on the feature space. The mean and standard deviation of bone lengths, grouped by participant, are shown in Table 5.

4.2. Automated human recognition

We demonstrate the effectiveness of our novel human recognition feature as a discriminative feature able to differentiate between individuals without humanly-recognisable data (i.e., colour images) using two common approaches: batch processing with the k-nearest neighbour (k-NN) algorithm and real-time processing with a multi-layer perceptron (MLP) model. While a plethora of classification algorithms exist (e.g., SVM Cortes & Vapnik, 1995, decision tree and random forest Ho, 1995) for these classification tasks, k-NN and MLP are highly prevalent in the literature and are suitable representatives for the evaluation of our new feature. We use k-fold ($k = 5$) to randomly split our dataset into training and validation sets. All experiments are run on the Compute Canada Graham cluster (2 CPUs, 16 GB memory and Tesla T4 GPU) (Compute Canada, 2022). On the cluster, our maximum runtime was limited to six hours, and the following modules (and dependencies) were used for the experiments: StdEnv/2020, gcc/9.3.0, cuda/11.0, OpenCV/4.4.0, and cudnn.

4.2.1. Batch processing human recognition: k-nearest neighbour

Several statistical classification approaches currently exist that have been shown to be effective with datasets like ours. Building on the work of Azarloo and Farokhi (2012), García et al. (2008), Kayikcioglu and Aydemir (2010) and Pacheco and López (2019), we consider the k-NN algorithm (Altman, 1992); a non-parametric method which is arguably the most naive and commonly used method for classification and regression (Altman, 1992). The k-NN algorithm has been directly applied to images in a broad range of applications such as handwriting recognition (Wu & Zhang, 2010) and medical image classification (Warfield, 1996). The k-NN algorithm is also effective as a classifier after data have been processed by other feature-extracting algorithms (Hodge, Lees, & Austin, 2004; Tico, Immonen, Ramo, Kuosmanen, & Saarinen, 2001). One of the most significant advantages of k-NN is that it is a relatively simple and straight-forward algorithm, but the classification accuracy is often comparable to more sophisticated approaches (Roli & Fumera, 2001). The speed of k-NN, however, worsens exponentially when the size of the training dataset increases (Hodge et al., 2004).

Our 20-dimensional bone segment feature vector is theoretically suitable for classification of individuals using k-NN (Pestov, 2013). Described in Algorithm 1, we classify an individual in a frame by finding the closest cluster the body segment feature vector belongs to.

Algorithm 1 Pseudo code for k-NN classification

```
procedure K-NN(training data, input)
    $d_{min} \leftarrow \infty$
    $label \leftarrow 0$
    $bones_{input} \leftarrow$ extract_bone_lengths(input)
    for frame in training data do
        $bones \leftarrow$ extract_bone_lengths(frame)
        $d \leftarrow$ euclidean_distance($bones_{input}, bones$)
        if $d < d_{min}$ then
            $d_{min} \leftarrow d$
            $label \leftarrow frame.label$
        end if
    end for
    return label
end procedure
```

Using the entire bone segment feature space dataset, we built a k-NN model, setting $k = 5$, determined using the Elbow method (Thorndike, 1953). Since k-NN does not involve a formal training step, and instead the training data are stored as the model, we evaluated the k-NN model accuracy by running k-fold cross-validation with a common choice of $k = 5$ (Anguita, Ghelardoni, Ghio, Oneto, & Ridella, 2012) on the training data, resulting in a model score of 0.96, found by taking the mean of the k-fold model scores. We then evaluated the performance of our model against the validation dataset individually for each participant per frame, as well as the overall performance of our classifier, which achieved an overall error rate of 3.37% across 1,447,560 frames (see Table 6). This error rate is for frame-by-frame, or per-skeleton classification across all participants.

Our initial frame-by-frame skeleton recognition is performed on each new frame of skeleton data without recognition of the temporal relationship between frames. The Microsoft Kinect SDK has a built-in feature that tracks instances of individuals across consecutive frames, assigning a globally unique, short-term *tracking ID* to each frame of

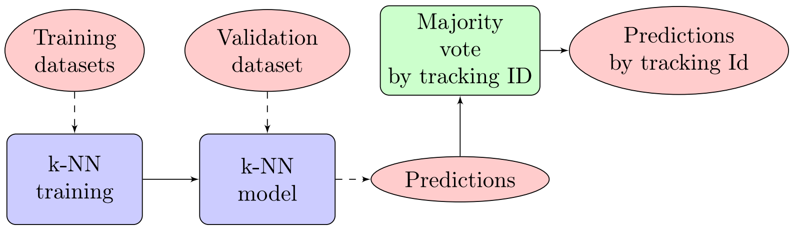

Fig. 3. Flowchart for k-NN-based batch processing instance-based human recognition.

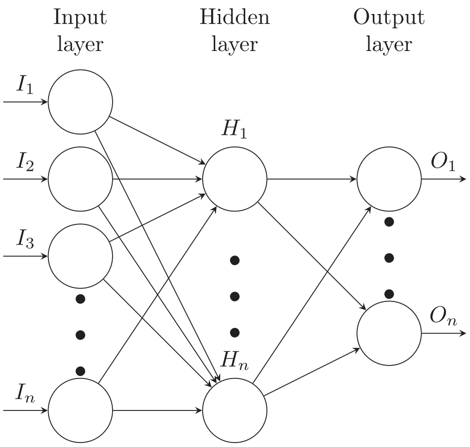

Fig. 4. Multilayer perceptron model for human recognition.

Table 4
Bone segment definition using joints provided by the Kinect SDK. Left- and right-side segments are denoted with _L and _R suffixes. SHOULDER is abbreviated SHLDR.

Bone	Segment label	SDK start joint	SDK end joint
0	SPINE	SPINE_BASE	SPINE_MID
1	SHLDR	SPINE_MID	SPINE_SHLDR
2	NECK	SPINE_SHLDR	NECK
3	HEAD	NECK	HEAD
4	SHLDR_L	SPINE_SHLDR	SPINE_SHLDR_L
5	ELBOW_L	SPINE_SHLDR_L	ELBOW_L
6	WRIST_L	ELBOW_L	WRIST_L
7	HAND_L	WRIST_L	HAND_L
8	SHLDR_R	SPINE_SHLDR	SPINE_SHLDR_R
9	ELBOW_R	SPINE_SHLDR_R	ELBOW_R
10	WRIST_R	ELBOW_R	WRIST_R
11	HAND_R	WRIST_R	HAND_R
12	HIP_L	SPINE_BASE	HIP_L
13	KNEE_L	HIP_L	KNEE_L
14	ANKLE_l	KNEE_L	ANKLE_L
15	FOOT_L	ANKLE_L	FOOT_L
16	HIP_R	SPINE_BASE	HIP_R
17	KNEE_R	HIP_R	KNEE_R
18	ANKLE_R	KNEE_R	ANKLE_R
19	FOOT_R	ANKLE_R	FOOT_R

skeleton data. Each instance of a tracked skeleton is assigned a new *tracking ID* when a skeleton's tracking is lost (e.g. an individual walks out of the field-of-view of the sensor and then back in, or an individual's physical shape deforms or is occluded beyond recognition as a human and then returns to a condition that can be tracked). As a result, a single person is generally assigned multiple *tracking IDs* throughout each single activity, and our use case prevents us from knowing *a priori* which unique *tracking IDs* belong to the same real person. That is to say, one real person will certainly be assigned different *tracking IDs* over an entire hour, and in general will be assigned multiple different *tracking IDs* even within a single activity. To improve on the limited instance *tracking ID* method performed by the SDK, we implemented a majority vote on each set of consecutive predicted participant IDs.

Our overall human participant batch-processing pipeline is illustrated in Fig. 3. Using this simple majority vote across consecutive frame-by-frame predictions, we observed that our k-NN model achieved 100% instance-based participant prediction accuracy across all participants.

Notably, batch processing models like k-NN commonly suffer from long processing times relative to real-time classifiers, albeit with higher classification accuracies in well-clustered datasets. Our batch processing k-NN implementation, for example, required 3010 s (50 m:10 s) to finish the k-fold evaluation. On average, it processes 975.1 samples per second. The processing speed reduced when the number of samples increased. Accordingly, the computational resource requirements of k-NN make it impractical for real-time human recognition. Still, the literal perfect classification results strongly support the discriminative nature of our novel skeleton feature, which makes k-NN ideal for accuracy-sensitive batch processing analysis.

4.2.2. *Real-time human recognition*

To evaluate the effectiveness of our skeleton feature in differentiating between individuals in real time, we implemented the popular multi-layer perceptron (MLP) model (Hastie, Tibshirani, & Friedman, 2009). In particular, considering the limitation of the k-NN for real-time applications we selected the MLP, which is a widely used artificial neural network for classification tasks that is capable of approximating arbitrarily complex functions (Csáji et al., 2001). MLP uses a training process to build an abstract model for the training data, and unlike k-NN, classification is much faster since the input sample is only computed with the (relatively) small model.

The structure of our three-layer MLP is described in Fig. 4. We use the Tensorflow framework (Abadi et al., 2016) to build, train and classify the bone segment dataset. The full structure of the network is described in Table 7.

We first evaluated the frame-by-frame performance of our MLP model before implementing the majority voting across consecutive frames to evaluate the model's instance-based human recognition performance. Our overall human participant real-time processing pipeline is illustrated in Fig. 5

Our MLP network, described in Fig. 4, had 1024 neurons in the hidden layer, with each hidden layer neuron fully connected to the input and output layers. We used the ReLu (Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000) activation function for the hidden layer because of it's fast convergence and sparsity (Qiu, Jiang, Pei, Lu, et al., 2017). Again, we used k-fold cross-validation with $k = 5$ on each frame with our MLP network which resulted in an average classification error rate of 0.1418 per skeleton. K-fold cross-validation took 9535 s (158 min and 55 s), including training and prediction. On average, prediction took 13.53 s, and the MLP model processed 289,511 samples per second. We evaluated the trained MLP model on the validation dataset, and we show the classification performance in Table 8. Following the same approach described in Section 4.2.1, we evaluated the instance-based classification accuracy (i.e., per tracking ID) of our MLP model. The averaged instance-based error rate over the 1203 independent instances was 0.034 (i.e., 96.57% accuracy), distributed relatively evenly across all participants.

In real-time recognition models we do not have access to all data at once. Instead, we collect a small amount of data for each participant, then apply the learned model to new data collected over the entire

Table 5

Mean (Standard deviation) of feature space for each participant (×10^{-2}).

00	01	02	03	04	05	06	07	08	09	10	11	12	13
Bone 00	8.74 (0.79)	12.1 (2.21)	10.5 (1.35)	9.54 (1.45)	9.88 (1.15)	7.13 (2.39)	8.96 (1.04)	9.58 (1.15)	9.77 (1.14)	10.61 (0.98)	9.51 (2.39)	8.88 (0.96)	18.49 (0.51)
Bone 01	4.65 (0.42)	6.42 (3.37)	5.57 (0.71)	5.19 (9.3)	5.29 (7.12)	3.84 (1.29)	4.81 (0.54)	5.16 (5.24)	5.19 (2.26)	5.73 (0.5)	5.05 (1.29)	4.85 (9.51)	9.86 (0.15)
Bone 02	0.49 (0.04)	0.67 (0.12)	0.59 (0.08)	0.54 (0.08)	0.56 (0.06)	0.41 (0.14)	0.51 (0.06)	0.54 (0.07)	0.55 (0.07)	0.61 (0.05)	0.53 (0.14)	0.51 (0.06)	1.04 (0.01)
Bone 03	1.86 (0.16)	2.21 (0.43)	2.04 (0.24)	1.88 (0.27)	1.5 (0.22)	1.61 (0.26)	1.88 (0.23)	2.08 (0.2)	2.04 (0.26)	2 (0.20)	1.31 (0.25)	2.27 (0.22)	2.55 (0.01)
Bone 04	3.35 (0.65)	4.51 (8.48)	3.76 (0.58)	3.7 (16.72)	3.59 (9.6)	2.99 (0.74)	2.83 (1.92)	3.19 (7.74)	3.22 (2.7)	3.51 (0.59)	3.07 (0.89)	2.88 (17.15)	2.09 (0)
Bone 05	6.23 (0.76)	7.26 (1.36)	7.88 (1.16)	8.15 (1.04)	5.8 (1.33)	4.06 (0.97)	6.19 (0.88)	7.12 (1.26)	6.9 (1.22)	6.53 (0.81)	5.41 (1.25)	4.38 (0.97)	7.97 (0.01)
Bone 06	4.64 (0.65)	7.29 (1.55)	5.82 (0.84)	5.66 (0.65)	4.1 (0.97)	4.01 (1.26)	4.43 (0.81)	4.97 (0.91)	5.72 (1.35)	5.59 (0.78)	4.93 (1.47)	5.1 (1.08)	6.48 (0)
Bone 07	0.21 (0.37)	0.53 (5.96)	0.49 (0.35)	0.31 (12.51)	0.2 (0.23)	0.75 (2.11)	0.16 (0.42)	0.17 (0.83)	0.52 (0.87)	0.49 (0.23)	0.67 (1.02)	0.46 (5.85)	0.78 (0)
Bone 08	3.15 (0.45)	4.96 (11.34)	3.61 (0.56)	3.85 (18.36)	3.32 (8.46)	2.78 (1.34)	3.1 (0.84)	3.13 (5.06)	3.11 (7.15)	3.33 (0.46)	3 (0.59)	2.97 (12.24)	1.02 (0.06)
Bone 09	5.94 (0.94)	8.26 (1.78)	7.95 (1.48)	7.28 (1.29)	5.47 (1.02)	4.87 (0.87)	6.17 (1.05)	6.35 (1.11)	6.63 (1.4)	6.85 (1.22)	5.63 (1.03)	5.07 (0.95)	8.16 (0.01)
Bone 10	4.82 (0.8)	6.62 (1.29)	6.15 (1.3)	5.86 (1.01)	4.29 (0.69)	4.19 (1.13)	4.47 (0.74)	4.92 (0.89)	5.79 (1.23)	5.51 (0.99)	4.37 (1.31)	4.38 (0.95)	3.55 (0.04)
Bone 11	0.39 (0.32)	0.77 (8.13)	0.56 (0.42)	0.57 (5.5)	0.22 (0.18)	0.76 (1.18)	0.31 (0.41)	0.29 (0.53)	0.7 (1.62)	0.67 (0.52)	0.49 (0.8)	0.56 (9.81)	0.58 (0.11)
Bone 12	0.71 (0.16)	0.98 (0.27)	0.74 (0.18)	0.78 (2.46)	0.56 (0.17)	0.59 (0.24)	0.54 (0.14)	0.59 (0.24)	0.59 (0.24)	0.81 (0.16)	0.84 (0.45)	0.58 (0.3)	0.31 (0.01)
Bone 13	11.37 (1.89)	14.52 (3.25)	19.66 (2.61)	17.21 (1.99)	10.31 (8.95)	12.37 (7.14)	12.04 (1.81)	13.28 (2.34)	15.88 (6.1)	14.99 (3.2)	11.19 (2.02)	11.97 (8.44)	29.35 (0.83)
Bone 14	7.46 (2.1)	12.30 (6.6)	10.51 (2.66)	10.9 (3.01)	5.28 (3.06)	8.11 (5.4)	7.49 (2.23)	7.46 (2.92)	9.36 (1.89)	10.29 (2.32)	8.63 (2.01)	9.21 (5.86)	12.54 (0)
Bone 15	1.33 (0.27)	1.80 (0.43)	1.51 (0.3)	1.32 (0.38)	1.17 (0.32)	1.36 (0.45)	1.37 (0.33)	1.51 (0.38)	1.51 (0.43)	1.46 (0.3)	1.24 (0.44)	1.33 (0.33)	0.68 (0.05)
Bone 16	0.73 (0.16)	0.96 (0.34)	0.88 (0.27)	0.86 (2.22)	0.77 (0.18)	0.68 (0.22)	0.71 (0.15)	0.77 (0.2)	0.78 (0.27)	0.79 (0.24)	0.94 (0.43)	0.63 (0.26)	0.27 (0.02)
Bone 17	11.06 (1.81)	15.02 (3.27)	19.99 (3.03)	17.69 (2.68)	10.06 (2.09)	11.03 (6.8)	12.79 (1.95)	13.17 (2.36)	16.48 (3.26)	14.69 (3.65)	11.39 (1.89)	11.96 (2.57)	33.33 (0.39)
Bone 18	7.94 (1.99)	13.36 (6.79)	10.63 (2.81)	10.96 (2.15)	5.52 (4.56)	8.38 (4.38)	7.25 (2.24)	7.37 (2.5)	9.14 (5.95)	10.52 (2.47)	10.25 (2.76)	9.59 (4.21)	16.37 (0)
Bone 19	1.29 (0.31)	1.46 (0.40)	1.48 (0.32)	1.60 (0.45)	1.23 (0.28)	1.52 (0.47)	1.31 (0.29)	1.45 (0.33)	1.53 (0.49)	1.30 (0.32)	1.49 (0.41)	1.23 (0.35)	2.21 (0.06)

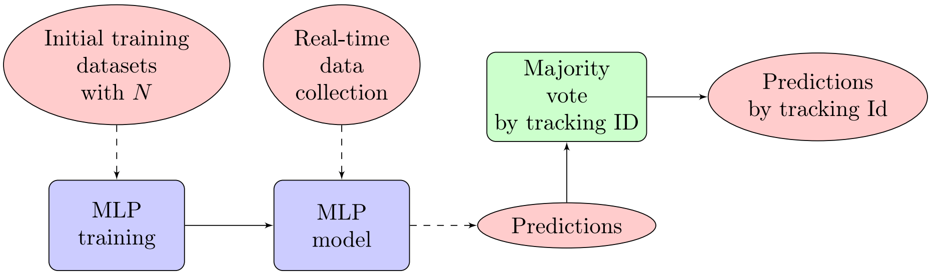
Initial training datasets with N
Real-time data collection
Majority vote by tracking ID
Predictions by tracking Id
MLP training
MLP model
Predictions

Fig. 5. Flowchart real-time human recognition.

Table 6
Results of k-NN model, averaged k-fold cross-validation.

Label	Frames	Error per frame voted	Rate per frame voted	Error per frame raw	Rate per frame raw	# tracking ids	Error per tracking id	Rate per tracking id
0	23 106	0	0	1084.8	0.046943607	99	0	0
1	39 301.8	0	0	1700.6	0.04326892	112	0	0
2	17 007	0	0	393.8	0.023151692	80	0	0
3	31 859	0	0	835	0.026209868	152	0	0
4	30 723	0	0	683.6	0.022249917	164	0	0
5	46 160.6	0	0	708.8	0.015354919	172	0	0
6	8641.2	0	0	142.6	0.016485888	19	0	0
7	7975	0	0	321.6	0.040317674	30	0	0
8	16 865.2	0	0	684.6	0.040592983	67	0	0
9	38 062.2	0	0	1517.6	0.039870432	125	0	0
10	9678.2	0	0	341.8	0.03529461	59	0	0
11	1745.4	0	0	108.2	0.062009302	6	0	0
12	18 387.2	0	0	501.6	0.027284115	118	0	0
Average	22 270.13846	0	0	694.2	0.033771841	92.53846154	0	0

Table 7
MLP structure.

Layer	Type	Number of neurons	Activation function
Input	Dense	20	ReLU
Hidden	Dense	1024	ReLU
Dropout	Dropout	N/A	N/A
Output	Dense	13	Softmax

session. The practical implication of this is that the quality of the training data may impact the overall classification accuracy. In order to evaluate the robustness of our skeleton as a feature for training our model, we created a simulated process where we randomly chose N frames from each participant that would comprise the training dataset, and then used these randomly sampled datasets to train MLP models. Finally, we evaluated the trained models against our full dataset. The results, visualised in Fig. 6, show that the MLP classification error rate significantly increases once $N < 2^9$. We chose N empirically, initially setting $N = 2^{13}$ and dividing our sample in half, ultimately evaluating training sample sizes where $N = (2^{13}, 2^{12} \dots 2^5, 2^4)$ (see Fig. 5).

5. Discussion and conclusions

We hypothesised that our novel 3D skeleton feature, comprised of bone segment lengths identified as the Euclidean distance between anatomical body parts found through semantic labelling, can be used to uniquely discriminate between different individual humans. Specifically, using these novel features, we implemented the common k-nearest neighbours (k-NN) and multi-layer perceptron (MLP) classifiers both on individual frames of 3D skeleton data (per-frame), and then implemented a simple majority vote (instance-based) aggregation across consecutive frames of data.

Overall, our k-NN classifier achieved a very high (3.37% error rate) per-frame accuracy and perfect instance-based classification with majority vote, at the cost of significant processing time (3010 s). This slow processing time severely restricts the use of k-NN in real-time applications. However, the high classification accuracy strongly supports our original hypothesis; that our novel 3D bone length features can effectively discriminate between different individuals over our labelled validation dataset of 1,447,559 frames of skeleton data (see Table 6). Our frame-by-frame MLP model also achieved a high classification accuracy of 85.81%, and instance-based accuracy of 96.57%. While lower than our k-NN model, these results are still objectively quite high and support the discriminative nature of our novel feature with real time processing. Compared to the results of Wang et al. (2003) which varied from 63.75% to 93.75%, our results are more reliable across classes, which varied from 79.70% to 97.40%. Notably, our approach of using 3D skeleton data is also theoretically independent of sensor perspective, compared to the 2D silhouette approach of Wang et al. (2003) which will be highly reliant on perspectives parallel to the ground plane.

Our approach, regardless of classifier, is entirely dependent on the successful acquisition of skeleton data. While our approach is robust to the noise inherent in both the depth data captured by the sensor and the joint inference performed by the Kinect SDK, we only included frames of data that had all 25 joints tracked. Furthermore, in some cases (e.g., where participants were seated on a chair or in a wheelchair) we may not be able to extract a full skeleton at any point in time. In the future, we will evaluate the performance of our classifier in conditions where only partial or unreliable skeleton data are available. Additionally, while our dataset was very large in terms of the number of image frames and the number of trials contributed by our participants, in the future we will evaluate our approach on a larger number of unique individuals.

Overall our results suggest that by using only depth data captured by an RGB-D sensor, we can accurately discriminate between individuals in group settings, including across multiple sessions (e.g., over multiple days). By only using depth data, we arguably help preserve an individual's identity, which supports the use of unobtrusive computer

Table 8
Results MLP model, averaged k-Fold cross-validation.

Label	Frames	Error per frame (voted)	Rate per frame (voted)	Confidence interval (voted)	Error per frame (raw)	Rate per frame (raw)	Confidence interval (raw)	# tracking ids	Error per tracking id	Rate per tracking id
0	23 106	3212.8	0.138929	0.00446	6754.4	0.292228	0.005864	99	5.6	0.056566
1	39 301.8	5195.4	0.132146	0.003348	11 742.6	0.298742	0.004525	112	7.2	0.064286
2	17 007	88.2	0.005178	0.001079	2031.2	0.119425	0.004874	80	0	0
3	31 859	642.4	0.020194	0.001545	3773.8	0.118518	0.003549	152	0.8	0.005263
4	30 723	789.6	0.025693	0.001769	3086	0.100432	0.003361	164	2.2	0.013415
5	46 160.6	1321.2	0.028616	0.001521	5224.4	0.113183	0.00289	172	2.2	0.012791
6	8641.2	0	0	0	856.2	0.09907	0.006299	19	0	0
7	7975	991.4	0.124159	0.007238	2930.8	0.367533	0.010582	30	1	0.033333
8	16 865.2	2020.6	0.119787	0.004901	4521.6	0.268094	0.006685	67	1.2	0.01791
9	38 062.2	2299.8	0.060451	0.002394	7134.4	0.18745	0.003921	125	3.6	0.0288
10	9678.2	637	0.065961	0.004945	2140.2	0.221203	0.008269	59	0.4	0.00678
11	1745.4	44.6	0.026659	0.007557	497.2	0.285224	0.021183	6	0	0
12	18 387.2	380.4	0.020731	0.002059	1701.6	0.092635	0.004191	118	3	0.025424
Average	22 270.13	1355.64	0.059115		4030.3384	0.197210		92.53	2.0923	0.0203

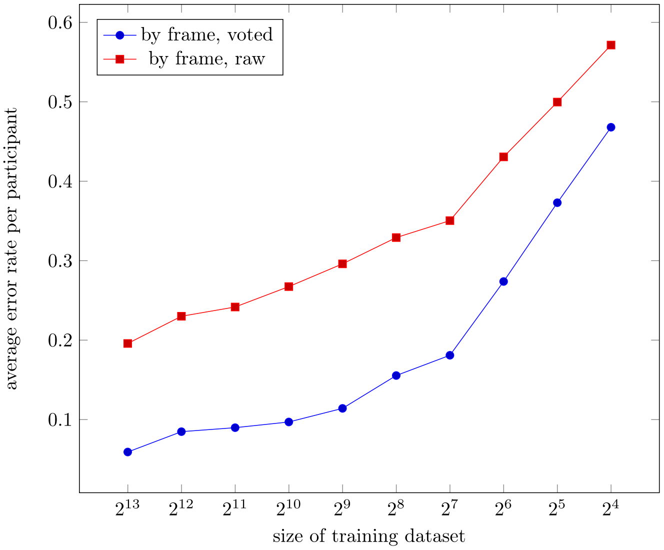

Fig. 6. Average error rate per tracking ID per participant, iteratively halving our training set size N from our initial value (2^{13}).

vision data in domains that are more sensitive, such as in applications involving children, nudity or private locations like bathrooms, potentially avoiding ethical issues and privacy concerns.

6. Limitations and future works

Our approach requires skeleton data as input, which is typically generated from depth images. This limits the applicability of our method to existing medical images, as the hardware required to collect depth images is not commonly available in most medical research settings, and depth data cannot be retroactively collected from existing images. Additionally, the bone-length feature we developed for frame-by-frame classification is purely spatial in nature, meaning that certain information (e.g., colour image data) may be lost during the de-identification process which could affect the future utility of the data. Additionally, temporal features such as the movements and gait of the skeleton between frames could be useful for improving the performance of the classification. Exploring the use of temporal features in future work could potentially enhance the performance of our method.

The batch processing approach we proposed has near perfect recognition accuracy, but can only be used in 'offline' scenarios where all of the data is available at the time of training. This makes it more suitable for 'proof-of-concept' demonstrations rather than real-world applications. Also, this approach can be useful for generating ground truth datasets that can be used to evaluate the performance of other methods.

It is also worth noting that although the skeletons we collected are anonymised and difficult for humans to recognise, certain physiological or soft-biometric features such as the behaviour of a series of skeletons could potentially be used to de-anonymise the participants.

CRediT authorship contribution statement

Zizui Chen: Data curation, Software, Validation, Visualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Stephen Czarnuch:** Funding acquisition, Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Erica Dove:** Data curation, Methodology, Investigation, Project administration, Writing – review & editing. **Arlene Astell:** Data curation, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Stephen Czarnuch reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. Stephen Czarnuch reports a relationship with Memorial University of Newfoundland that includes: employment.

Data availability

The data that have been used are confidential.

Acknowledgement

This work was supported by NSERC Discovery Grant RGPIN-2916-04165.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). TensorFlow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265–283). URL https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The'K'in K-fold cross validation. In *ESANN*.

Azarloo, A., & Farokhi, F. (2012). Automatic musical instrument recognition using K-NN and MLP neural networks. In *2012 fourth international conference on computational intelligence, communication systems and networks* (pp. 289–294).

Barmpoutis, A. (2013). Tensor body: Real-time reconstruction of the human body and avatar synthesis from RGB-D. *IEEE Transactions on Cybernetics*, *43*(5), 1347–1356. http://dx.doi.org/10.1109/TCYB.2013.2276430.

Basak, H., Kundu, R., Singh, P. K., Ijaz, M. F., Woźniak, M., & Sarkar, R. (2022). A union of deep learning and swarm-based optimization for 3D human action recognition. *Scientific Reports*, *12*(1), 1–17.

Chen, Z. (2020). Skeleton visualiser. https://github.com/bearnl/skeleton-visualiser.

Chen, H.-S., Chen, H.-T., Chen, Y.-W., & Lee, S.-Y. (2006). Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on video surveillance and sensor networks* VSSN '06, (pp. 171–178). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1178782.1178808.

Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., & Yu, N. (2017). Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In *Proceedings of the IEEE international conference on computer vision* (pp. 4836–4845).

Cippitelli, E., Gasparrini, S., Gambi, E., & Spinsante, S. (2016). A human activity recognition system using skeleton data from rgbd sensors. *Computational Intelligence and Neuroscience*, *2016*, 21.

(2022). [link]. URL https://www.computecanada.ca/advanced-research-computing/national-systems/.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Csáji, B. C., et al. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, *24*(48), 7.

Czarnuch, S., & Mihailidis, A. (2016). Development and evaluation of a hand tracker using depth images captured from an overhead perspective. *Disability and Rehabilitation: Assistive Technology*, *11*(2), 150–157.

Ding, C., & Tao, D. (2015). Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, *17*(11), 2049–2058.

Dove, E., & Astell, A. (2019). The Kinect Project: Group motion-based gaming for people living with dementia. *Dementia*, *18*(6), 2189–2205.

Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1110–1118).

Fan, J., Xu, W., Wu, Y., & Gong, Y. (2010). Human tracking using convolutional neural networks. *IEEE Transactions on Neural Networks*, *21*(10), 1610–1623.

Feng, D., Haase-Schuetz, C., Rosenbaum, L., Hertlein, H., Duffhauss, F., Glaeser, C., et al. (2019). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. arXiv preprint arXiv:1902.07830.

Franco, A., Magnani, A., & Maio, D. (2020). A multimodal approach for human activity recognition based on skeleton and RGB data. *Pattern Recognition Letters*, *131*, 293–299. http://dx.doi.org/10.1016/j.patrec.2020.01.010, URL https://www.sciencedirect.com/science/article/pii/S0167865520300106.

Gaglio, S., Re, G. L., & Morana, M. (2015). Human activity recognition process using 3-D posture data. *IEEE Transactions on Human-Machine Systems*, *45*(5), 586–597. http://dx.doi.org/10.1109/THMS.2014.2377111.

García, V., Mollineda, R. A., & Sánchez, J. S. (2008). On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, *11*(3–4), 269–280.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857.

Gonzalez, R. C., & Woods, R. E. (2006). Digital image processing (3rd edition). USA: Prentice-Hall, Inc., ISBN: 013168728X.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*(6789), 947–951.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition, Vol. 1* (pp. 278–282). IEEE.

Hodge, V. J., Lees, K. J., & Austin, J. L. (2004). A high performance k-NN approach using binary neural networks. *Neural Networks*, *17*(3), 441–458.

Hoey, J., Poupart, P., von Bertoldi, A., Craig, T., Boutilier, C., & Mihailidis, A. (2010). Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. *Computer Vision and Image Understanding*, *114*(5), 503–519.

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2012). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *NIPS*.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *14*(1), 4–20.

Jiang, H., & Learned-Miller, E. (2017). Face detection with the faster R-CNN. In *2017 12th IEEE international conference on automatic face gesture recognition (FG 2017)* (pp. 650–657). http://dx.doi.org/10.1109/FG.2017.82.

Kayikcioglu, T., & Aydemir, O. (2010). A polynomial fitting and k-NN based approach for improving classification of motor imagery BCI data. *Pattern Recognition Letters*, *31*(11), 1207–1215.

Khan, M. Z., Harous, S., Hassan, S. U., Ghani Khan, M. U., Iqbal, R., & Mumtaz, S. (2019). Deep unified model for face recognition based on convolution neural network and edge computing. *IEEE Access*, *7*, 72622–72633. http://dx.doi.org/10.1109/ACCESS.2019.2918275.

Kim, H.-J., Lee, J. S., & Park, J. (2008). Dynamic hand gesture recognition using a CNN model with 3D receptive fields. In *2008 international conference on neural networks and signal processing* (pp. 14–19). http://dx.doi.org/10.1109/ICNNSP.2008.4590300.

Koo, J. H., Cho, S. W., Baek, N. R., Kim, M. C., & Park, K. R. (2018). CNN-based multimodal human recognition in surveillance environments. *Sensors*, *18*(9), 3040.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., et al. (2016). Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2176–2184).

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113.

Le, T.-L., Nguyen, M.-Q., et al. (2013). Human posture recognition using human skeleton provided by Kinect. In *2013 international conference on computing, management and telecommunications (ComManTel)* (pp. 340–345). IEEE.

Li, C., Hou, Y., Wang, P., & Li, W. (2017). Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, *24*(5), 624–628. http://dx.doi.org/10.1109/lsp.2017.2678539.

Liang, X., Wei, Y., Lin, L., Chen, Y., Shen, X., Yang, J., et al. (2016). Learning to segment human by watching youtube. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(7), 1462–1468.

Liu, J., Akhtar, N., & Mian, A. (2017). Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition. arXiv:1711.05941.

Liu, M., Liu, H., & Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, *68*, 346–362. http://dx.doi.org/10.1016/j.patcog.2017.02.030, URL https://www.sciencedirect.com/science/article/pii/S0031320317300936.

Liu, J., Wang, G., Duan, L.-Y., Abdiyeva, K., & Kot, A. C. (2018). Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Transactions on Image Processing*, *27*(4), 1586–1599. http://dx.doi.org/10.1109/tip.2017.2785279.

Lu, Z., Jiang, X., & Kot, A. (2018). Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, *25*(4), 526–530.

Maltoni, D., Maio, D., Jain, A. K., & Prabhakar, S. (2009). *Handbook of fingerprint recognition*. Springer Science & Business Media.

Mesejo, P., Martos, R., Ibanez, O., Novo, J., & Ortega, M. (2020). A survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification. *Applied Sciences*, *10*(14), http://dx.doi.org/10.3390/app10144703, URL https://www.mdpi.com/2076-3417/10/14/4703.

Microsoft Kinect Developer (2017). [link]. URL https://developer.microsoft.com/en-us/windows/kinect/develop.

Nakajima, C., Pontil, M., Heisele, B., & Poggio, T. (2003). Full-body person recognition system. *Pattern Recognition*, *36*(9), 1997–2006.

Ng, K.-D., Mehdizadeh, S., Iaboni, A., Mansfield, A., Flint, A., & Taati, B. (2020). Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia. *IEEE Journal of Translational Engineering in Health and Medicine*, *8*, 1–9.

Pacheco, W. D. N., & López, F. R. J. (2019). Tomato classification according to organoleptic maturity (coloration) using machine learning algorithms K-NN, MLP, and K-Means Clustering. In *2019 XXII symposium on image, signal processing and artificial vision* STSIVA, (pp. 1–5).

Parajuli, M., Tran, D., Ma, W., & Sharma, D. (2012). Senior health monitoring using Kinect. In *2012 fourth international conference on communications and electronics* ICCE, (pp. 309–312). IEEE.

Pestov, V. (2013). Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, *65*(10), 1427–1437.

Preis, J., Kessel, M., Werner, M., & Linnhoff-Popien, C. (2012). Gait recognition with kinect. In *1st international workshop on kinect in pervasive computing* (pp. 1–4). UK: New Castle.

Qiu, S., Jiang, M.-y., Pei, Z.-l., Lu, Y.-n., et al. (2017). Text classification based on ReLU activation function of SAE algorithm. In *International symposium on neural networks* (pp. 44–50). Springer.

Rathgeb, C., Dantcheva, A., & Busch, C. (2019). Impact and detection of facial beautification in face recognition: an overview. *IEEE Access*, *7*, 152667–152678. http://dx.doi.org/10.1109/ACCESS.2019.2948526.

Roli, F., & Fumera, G. (2001). Support vector machines for remote sensing image classification. In *Image and signal processing for remote sensing VI, Vol. 4170* (pp. 160–166). International Society for Optics and Photonics.

Saha, S., Pal, M., Konar, A., & Janarthanan, R. (2013). Neural network based gesture recognition for elderly health care using kinect sensor. In *International conference on swarm, evolutionary, and memetic computing* (pp. 376–386). Springer.

Sapinski, T., Kaminska, D., Pelikant, A., & Anbarjafari, G. (2019). Emotion recognition from skeletal movements. *Entropy*, *21*(7), http://dx.doi.org/10.3390/e21070646, URL https://www.mdpi.com/1099-4300/21/7/646.

Shan, J., & Akella, S. (2014). 3D human action segmentation and recognition using pose kinetic energy. In *2014 IEEE international workshop on advanced robotics and its social impacts* (pp. 69–75). http://dx.doi.org/10.1109/ARSO.2014.7020983.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., et al. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, *56*(1), 116–124. http://dx.doi.org/10.1145/2398356.2398381.

Sinha, A., Chakravarty, K., Bhowmick, B., et al. (2013). Person identification using skeleton information from kinect. In *Proc. intl. conf. on advances in computer-human interactions* (pp. 101–108).

http://arxiv.org/abs/1502.00873 Sun, Y., Liang, D., Wang, X., & Tang, X. (2015). DeepID3: Face recognition with very deep neural networks. *CoRR*, *abs/1502.00873*, arXiv:1502.00873.

Sun, J., Wang, Y., Li, J., Wan, W., Cheng, D., & Zhang, H. (2018). View-invariant gait recognition based on kinect skeleton feature. *Multimedia Tools and Applications*, *77*(19), 24909–24935.

Teipel, S., König, A., Hoey, J., Kaye, J., Krüger, F., Robillard, J. M., et al. (2018). Use of nonintrusive sensor-based information and communication technology for real-world evidence for clinical trials in dementia. *Alzheimer's & Dementia*, *14*(9), 1216–1231.

Thorndike, R. L. (1953). Who belongs in the family. In *Psychometrika*. Citeseer.

Tico, M., Immonen, E., Ramo, P., Kuosmanen, P., & Saarinen, J. (2001). Fingerprint recognition using wavelet features. In *ISCAS 2001. the 2001 IEEE international symposium on circuits and systems (Cat. No. 01CH37196), Vol. 2* (pp. 21–24). IEEE.

Vizilter, Y., Gorbatsevich, V., Vorotnikov, A., & Kostromov, N. (2016). Real-time face identification via CNN and boosted hashing forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 78–86).

Wang, L., Tan, T., Ning, H., & Hu, W. (2003). Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(12), 1505–1518.

Wang, L., Xu, Y., Cheng, J., Xia, H., Yin, J., & Wu, J. (2018). Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE Access*, *6*, 17913–17922. http://dx.doi.org/10.1109/ACCESS.2018.2817253.

Warfield, S. (1996). Fast k-NN classification for multichannel image data. *Pattern Recognition Letters*, *17*(7), 713–721.

Wieczorek, M., Siłka, J., Woźniak, M., Garg, S., & Hassan, M. M. (2022). Lightweight convolutional neural network model for human face detection in risk situations. *IEEE Transactions on Industrial Informatics*, *18*(7), 4820–4829. http://dx.doi.org/10.1109/TII.2021.3129629.

Wu, M., & Zhang, Z. (2010). Handwritten digit classification using the mnist data set. In *Course project CSE802: Pattern classification & analysis*.

Xia, L., Chen, C.-C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3D joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 20–27). http://dx.doi.org/10.1109/CVPRW.2012.6239233.

Xiang, Y., Schmidt, T., Narayanan, V., & Fox, D. (2017). PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. arXiv:1711.00199.

Xiang, C., Zhang, L., Tang, Y., Zou, W., & Xu, C. (2018). MS-CapsNet: A novel multi-scale capsule network. *IEEE Signal Processing Letters*, *25*(12), 1850–1854.

Yao, L., Kusakunniran, W., Wu, Q., Zhang, J., Tang, Z., & Yang, W. (2021). Robust gait recognition using hybrid descriptors based on Skeleton Gait Energy Image. *Pattern Recognition Letters*, *150*, 289–296. http://dx.doi.org/10.1016/j.patrec.2019.05.012, URL https://www.sciencedirect.com/science/article/pii/S0167865519301618.

Zhang, C., & Czarnuch, S. (2020). Perspective independent ground plane estimation by 2D and 3D data analysis. *IEEE Access*, *8*(1), 82024–82034. http://dx.doi.org/10.1109/ACCESS.2020.2991346.

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE international conference on computer vision*. ICCV.

Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., et al. (2019). Pose2seg: detection free human instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 889–898).

Zhang, C., & Tian, Y. (2012). RGB-D camera-based daily living activity recognition. *Journal of Computer Vision and Image Processing*, *2*(4), 12.

Zhao, B., Feng, J., Wu, X., & Yan, S. (2017). A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, *14*(2), 119–135.

Zhao, Y., Liu, Z., Yang, L., & Cheng, H. (2012). Combing RGB and depth map features for human activity recognition. In *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference* (pp. 1–4).