entropy

MDPI

Article

Action Recognition Using Single-Pixel Time-of-Flight Detection

Ikechukwu Ofodile [1,†], Ahmed Helmi [1,†], Albert Clapés [2,†], Egils Avots [1,†], Kerttu Maria Peensoo [3,†], Sandhra-Mirella Valdma [3,†], Andreas Valdmann [3,†], Heli Valtna-Lukner [3,†], Sergey Omelkov [3,†], Sergio Escalera [2,4,†] and Cagri Ozcinar [5,†] and Gholamreza Anbarjafari [1,6,7,*,†]

1 iCv Lab, Institute of Technology, University of Tartu, 50411 Tartu, Estonia; ike@icv.tuit.ut.ee (I.O.); ahmed@icv.tuit.ut.ee (A.H.); ea@icv.tuit.ut.ee (E.A.)
2 University of Barcelona, 08007 Barcelona, Spain; aclapes@cvc.uab.es (A.C.); sergio@maia.ub.es (S.E.)
3 Institute of Physics, University of Tartu, 50411 Tartu, Estonia; kerttumariapeensoo@gmail.com (K.M.P.); sandhra91@gmail.com (S.-M.V.); andreas.valdmann@gmail.com (A.V.); heli.lukner@ut.ee (H.V.-L.); sergey.omelkov@ut.ee (S.O.)
4 The Computer Vision Centre, 08193 Barcelona, Spain
5 Trinity College Dublin, Dublin 2, Ireland; ozcinarc@scss.tcd.ie
6 Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep 27000, Turkey
7 Institute of Digital Technologies, Loughborough University London, London E15 2GZ, UK
* Correspondence: shb@icv.tuit.ut.ee; Tel.: +372-737-4855
† All authors contributed equally to this work.

Received: 14 January 2019; Accepted: 15 April 2019; Published: 18 April 2019

Abstract: Action recognition is a challenging task that plays an important role in many robotic systems, which highly depend on visual input feeds. However, due to privacy concerns, it is important to find a method which can recognise actions without using visual feed. In this paper, we propose a concept for detecting actions while preserving the test subject's privacy. Our proposed method relies only on recording the temporal evolution of light pulses scattered back from the scene. Such data trace to record one action contains a sequence of one-dimensional arrays of voltage values acquired by a single-pixel detector at 1 GHz repetition rate. Information about both the distance to the object and its shape are embedded in the traces. We apply machine learning in the form of recurrent neural networks for data analysis and demonstrate successful action recognition. The experimental results show that our proposed method could achieve on average 96.47% accuracy on the actions walking forward, walking backwards, sitting down, standing up and waving hand, using recurrent neural network.

Keywords: single pixel single photon image acquisition; time-of-flight; action recognition

1. Introduction

Action is a spatiotemporal sequence of patterns [1–6]. The ability to detect movement and recognise human actions and gestures would enable advanced human to machine interaction in wide scope of novel applications in the field of robotics from autonomous vehicles, surveillance for security or care-taking to entertainment.

In the field of machine vision, the majority of effort has been put into recognising human action from video sequences [7–9], because overwhelmingly imaging devices mimic the human-like perception of the surroundings and video format is most widely available. Videos are a sequence of two-dimensional intensity patters, captured by using an imaging lens projecting the scene to a two-dimensional detector array (a charge coupled device (CCD) device, for example). Unlike living

creatures, the ever growing field of robotics has run into major difficulties while trying to recognise objects, their actions, and their distances from two-dimensional images. Processing the data is computationally demanding, and depth information is not unanimously retrievable.

Deep neural networks, due to their high accuracy, are widely used in many of the computer vision applications such as emotion recognition [10–16], biometric recognition [17–20], personality analysis [21,22], and activity analysis [5,23,24]. Depending on the nature of the data, different structures can be used [25,26]. In this work, we deal with time-series data, i.e., we handle temporal information. For this purpose, we are mainly focused on recurrent neural networks (RNN) and long-short term (LSTM) algorithms.

In addition to colour and intensity, incident light can be characterised by its propagation direction(s), spectral content and temporal evolution in the case of pulsed illumination. Light also carries information about its source, and each medium, refraction, reflection and scattering event it has encountered or traversed. This enables various uncommon ways to characterise the scene. The rapid advancements in optoelectronics and availability of sufficient computational power enable innovative imaging and light capturing concepts, which serve as the ground for action detection. For example, a detector, capable of registering evolution of backscattered light with a high temporal resolution in a wide dynamic range, would be able to detect even objects hidden from the direct line of sight [27–29]. Along the same vein, several alternative light-based methods have been developed for resolving depth information or 3D map of the surroundings (some examples can be found in [30–34]) giving 3D information in voxel format about the scene, which is also suitable for action detection [35]. Combining the fundamental understanding of light propagation and computational neural networks for the data reconstruction, it appears that objects or even persons can be detected using a single pixel detector registering temporal evolution of the back-scattered light pulse [36].

In this work, which is a feasibility study of a novel setup and methodology for conducting action recognition, we propose and demonstrate an action recognition scheme based on a single-pixel direct time-of-flight detection. We use NAO robots in a controlled environment as a test subject. We illuminate a scene with a diverging picosecond laser pulse, (30 ps duration) and detect the temporal evolution of back scattered light with a single pixel multiphoton detector of 600 ps temporal resolution. Our data contains one-dimensional time sequences presenting the signal strength (proportional to the number of detected photons) versus arrival time. Information about both the distance to the object and its shape are embedded in the traces. We apply machine learning in the form of recurrent neural networks for data analysis and demonstrate successful action recognition.

The following list summarises the contributions of our work:

- Introduce an unexplored data modality for action recognition scenarios. In contrast to other depth-based modalities, our single-pixel light pulses are not visually interpretable which makes it a more privacy preserving solution.
- Provide a manually annotated dataset of 550 robot action sequences, some of them containing obstacle objects.
- Apply multi-layer bi-directional recurrent neural networks for the recognition task as an initial machine learning benchmarking baseline.
- Present an extensive set of experiments on the recognition of several action classes. We demonstrate how the learning models are able to extract proper action features and generalise several action concepts from captured data.

The rest of the paper is organised as follows: in Section 2, related works to single-pixel, single-photon acquisition and action recognition are reviewed. Section 3 describes the data collection and the details of the setup. In Section 4, the details of the proposed deep neural network algorithm used for action recognition are described. The experimental results and discussions are provided in Section 5. Finally, the work is concluded in Section 6.

2. Related Work

The filed of motion analysis was firstly inspired by intensity images and progresses towards depth images, which are more robust in comparison to intensity images. In the case of action recognition, the most useful are sensors that provide depth map. Nevertheless, data are processed to extract human silhouettes, body parts, skeleton and pose of the person, which in turn are used as features for machine learning methods to classify actions. These sensors have drawn much interest for human activity related research and software development.

2.1. Depth Sensors

Depth images provide the 3D structure of the scene, and can significantly simplify tasks such as background subtraction, segmentation, and motion estimation. With the recent advances in depth sensor hardware, such as time-of-flight (ToF) cameras, research based on depth imagery has appeared. Three main depth sensing technologies are applied in computer vision research: stereo cameras, time-of-flight (ToF) cameras and structured light.

1. Stereo cameras infer the 3D structure of a scene from two images from different viewpoints. The depth map is created using information about the camera setup (stereo triangulation) [37].
2. A time-of-flight (ToF) camera estimates distance to an object surface using active light pulses from a single camera, whose time to reflect from the object give the distance. Such devices use a sinusoidally modulated infra-red light signal, and distance is estimated using the phase shift of the reflected signal on CMOS or CCD detector. The most commercially know device that uses this technology is Kinect 2 [38,39], which provides depth map of 512 $\times$ 424 pixels at 30 frames per second.
3. Structured light sensors [40] such as Kinect 1 [41] which was released in November 2010 by Microsoft. Kinect 1 consists of an RGB camera and a depth sensor. The depth sensor provides depth map of 320 $\times$ 240 pixels at 30 frames per second.

Similar to ToF depth cameras, action can be encoded in a laser pulse, which is captured by single-pixel cameras. The contents of the scene are encoded in time-series data. When using single pixel camera setups, processing steps such as pose estimation is not necessary. The acquired time series data are usable for machine learning tasks without any modification or additional processing.

2.2. Sing-Pixel Single-Photon Acquisition

Recent advances in photonics offer various innovative approaches for three-dimensional imaging [42]. Among those is time-of-flight imaging, which enables detection and tracking of objects. This involves illuminating the scene with diverging light pulses shorter than 100 ps. The light is scattered back from the scene and its flight time is detected with respective accuracy. Flight time t of light multiplied by the speed c of light directly gives the distance the light pulse has travelled from the source to the detector. Often, the laser source and detector are nearby and the value ct equals twice the distance of the object. Compared to time-of-flight ranging used in LiDARs (Light Detection And Ranging device), the principles introduced here utilise the knowledge of light propagation and are potentially capable of achieving higher spatial resolution.

In early experiments, 50 fs pulse duration mode-locked Ti:Sapphire near-infrared (NIR) laser and streak camera of 15 ps temporal resolution with array matrix were used to detect movement in occluded environment or to recover the 3D shape of an object behind direct line of sight [28,43]. (Using light pulses as short as 50 fs was not necessary; this is a widely spread ultrashort pulse laser source available in photonics labs.) The reconstruction of the object shape required data traces from various viewing angles and mathematical back-projection. In the scope of current research, the non-line-of-sight illumination can be seen as a method of efficiently diverging the incident laser pulse on the scene. There has been several suggestions to use more widely accessible hardware by

replacing expensive and fragile streak camera with single photon avalanche diode (SPAD) [29], or to construct a setup based on modulated laser diodes and single pixel photonic mixer device [44].

In proof-of-principle experiment [45] a single pixel SPAD detector (the actual 32 × 32 pixels were used for statistics and to speed up the measurements, such device was and early prototype at the time) was used and ca. 50 ps temporal resolution was utilised to demonstrate the ability to detect linear movement of a non-line-of-sight object. Again, ultrashort 10 fs pulse duration Ti:Sapphire laser with carrier wavelength in NIR region was used. Instead of recording the shape of the object, the shape of its reflection on a screen (a floor) was recorded and position of the object was derived from geometry. Replacing the detector array by three single-pixel SPAD detectors, real-time movement of an object was traced [46]. In this experiment, pulsed NIR diode was used instead of Ti:Sapphire laser. The integration time for single-photon detector was reduced from approximately 3 s to 1 s. In consequent papers, the table-top scenes are scaled up to detect a human [36,47]. Significance of the solution presented in [36] relies on artificial neural network machine learning algorithms for data analysis instead of deterministic tools used before. As a result, the team led by Daniele Faccio was able to distinguish between several standing position of a human and distinct between three different persons by analysing merely one-dimensional trace of SPAD detector.

2.3. Action Recognition

Most action recognition and monitoring systems use images with high enough quality where a person can be identified. When considering commercial applications, such systems invade human privacy. The identification factor can be removed by blurring or obscuring the images, downscaling, using encryption and IT solutions to keep the stored data safe. Nevertheless, at some point data is available in a format where people can be identified and can be mishandled due to breach of security, selling private data for commercial purposes or by request from governmental authorities.

One way of removing the privacy concerns is to use devices which by default use low resolution images, hence eliminating privacy issues at the data acquisition step. For such purpose, researchers are developing methods for action recognition using single pixel and low-resolution cameras. A privacy preserving method was proposed Jia and Radke [48] to track a person and estimate pose of a person using a network of ceiling-mounted time-of-flight sensors. Tao et al. [49] based their solution on a network of ceiling-mounted binary passive infrared sensors to recognise a set of daily activities. Kawashima et al. [50] used extremely low-resolution (16 × 16 pixels) infrared sensors to monitor a person constantly day and night without privacy concerns. Ji Dai et al. [51] studied the privacy implications using virtual space for action recognition. They studied Kinect 2 resolutions from 100 × 100 pixels down to 1 × 1 and their effect on action recognition methods. To address privacy issues, Xu et al. [52] proposed a fully-coupled two-stream spatiotemporal architecture for reliable human action recognition on extremely low resolution (e.g., 12 × 16 pixel) videos.

In this research work, we develop a new methodology for action recognition without using any data which can rise a privacy issue. Such a system can be highly used in places such as nursery and hospitals where recognition of actions might be important without violating the privacy rights of people in the environment.

2.4. Data Interpretability

In comparison to devices such as Kinect, the depth map provides enough information about a person's body shape and height, and facial features to visually identify the person and his/her actions in the scene. In the proposed experimental setup, we recorded a kind of a depth map, but it was recorded with a single pixel detector. Hence, the trace has no spatial resolution, which would enable identifying a person or an object directly through detecting above mentioned properties. The spatial properties of the scene are imprinted into the temporal evolution of the recorded trace. In the case an action takes place, characteristic temporal evolution pattern is imprinted to the recorded trace. The recorded 1D time series containing temporal evolution of back scattered light (timestamped

detected photon amplitudes) is enough to recognise human actions when interpreted using machine learning algorithms. In the case of a static scene, there is no change in the consequent temporal traces, indicating that no actions are taking place. In addition, the data footprint of a 1D data trace is smaller than that of a depth map. This enables rapid processing times. In the case of using Kinect, the data processing pipeline contains human interpretable data that could be used for unlawful purposes, but in the proposed setup such possibility does not exist.

3. Collected Data

In this research, for data collection, we created a special setup. Figure 1 shows the general data collection setup, including the placement of the laser and the detector sensor. The data has been collected under the control environment where a NAO V4 humanoid robot was placed in a black box with dimensions of $800 \times 800 \times 1200$ mm^3 (W × H × L) and was used to conduct some pre-defined actions. The scene was illuminated by Fianium supecontinuum laser source (SC400-2-PP) working at 1 MHz rate. The scatterer ensured that the whole scene Was illuminated at once, without any scanning or other moving parts required. The reflected light from the scene was collected by a Hamamatsu R10467U-06 hybrid photodetector (HPD) with spectral sensitivity range of 220–650 nm. The neutral density filter (OD2) was used in front of the detector to prevent HPD damage due to overexposure. The signal from the HPD was boosted by a Hamamatsu C10778 preamplifier (37 dB, inverting) and then directly digitised by LeCroy WaveRunner 6100a (1 GHz, 10 Gs/s) oscilloscope. The HPD was used in a pulse current (multiphoton) mode, therefore the time resolution of the system was determined by its single-photon pulse response of 600 ps FWHM. The oscilloscope worked in a sequence acquisition mode, recording 200 traces from subsequent trigger events during one sequence, with average frame rate of five sequences per second. The traces within one sequence were averaged to improve signal-to-noise ratio due to both electronic noise and photon statistics. The usage of multiphoton detection mode allowed greatly reducing acquisition time per frame, although with a lower time resolution, unlike the single photon detection used in [46]. The oscilloscope traces in the form of reflected light intensity versus time in nanosecond scale contained all the relevant information about the scene in a non-human-readable form, thus preserving privacy. The series of such traces recorded a 5 fps therefore contain the information about motion.

Various experiments were performed using one- and two-robot setups. A short summary can be seen in Table 1, More detailed description of the tasks can be found in the following sections.

Table 1. Summary of the performed actions.

	One-Robot						Two-Robot	
Task	**Walk Forward**	**Walk Reverse**	**Sit Down**	**Stand Up**	**Hand Wave**	**Object Setup**	**Same Action**	**Different Actions**
Repetitions	125	125	50	50	50	156	70	20

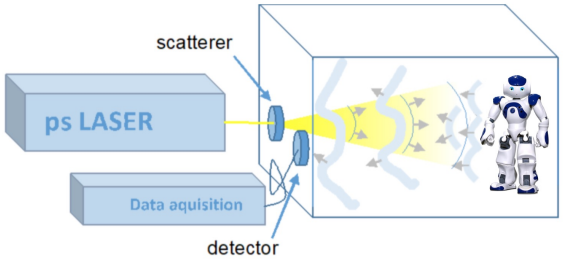


Figure 1. The data collection setup: Fianium laser delivers 30 ps duration light pulses. The collimated laser beam is directed to a scatterer, which creates divergent speckle pattern (giving divergence of 40 degree apex angle) inside the box, which are directed to the black box specially designed for the robot. Scattering illumination will reduce potential interference effects at the detector and, using controlled speckle pattern could be used to increase the lateral resolution. The light scattered from the moving object (NAO V4) and the walls is detected using single-pixel hybrid photodetector (HPD), which detects the temporal evolution of back scattered light.

3.1. ONE-Robot Setup

Initially, for acquiring training data, only one robot was used. Experiments were divided into the following categories:

1. **Directional walk**: We specified three starting points (A, B, and C) and three end points (a, b, and c) inside the box. The robot walked from starting points at 70 cm distance to corresponding end points, and vice versa. In addition, two diagonal directions, from Point A to Point c and from Point C to Point a, were travelled both forward and reverse, as illustrated in Figure 2. All action were repeated 25 times for each point per each direction. These walking actions are shown in Tables 2 and 3.
2. **Sitting down (sd) from standing up pose and Standing up (su) from sitting down pose**: We specified five areas where the robot was located, as illustrated in Figure 3. These actions were repeated 10 times per area and in each repetition, the position of the robot was nearly the same. Summary of performed tasks is shown in Table 4.
3. **Waving right hand (hw) for 3 s**: This action was repeated 25 times in Areas 2 and 5 (see Table 4).
4. **Include both object and robot**: An object was placed in the environment while the robot was doing the six tasks, which are listed in Table 5 (see Figure 4). Each task was repeated 12 times.

Table 2. Forward (F) movement.

Task	A1	A2	B1	C1	C2
Repetitions	25	25	25	25	25
Start location	A	A	B	C	C
Stop location	a	c	b	c	a

Table 3. Reverse (R) movement.

Task	A1	A2	B1	C1	C2
Repetitions	25	25	25	25	25
Start location	a	c	b	c	a
Stop location	A	A	B	C	C

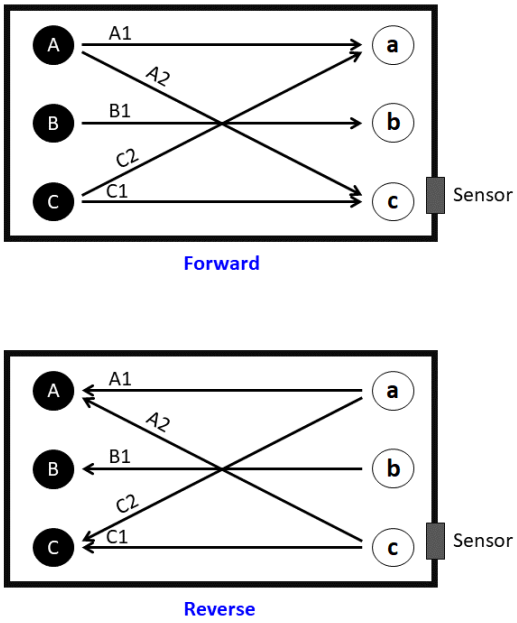


Figure 2. Start and endpoints, showing paths of the robot during directional walk.

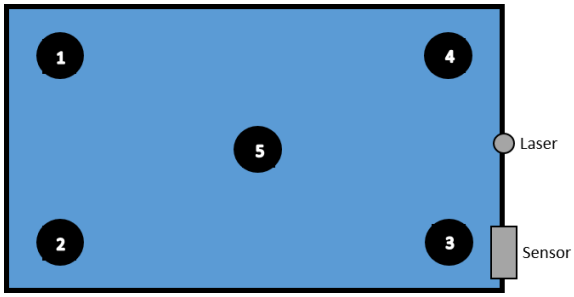


Figure 3. Positions of sitting down and standing up actions.

Table 4. Tasks performed in specific locations.

Task	**Sit Down**							**Stand Up**			**Hand-Wave**	
Repetitions	10	10	10	10	10	10	10	10	10	10	25	25
Location	1	2	3	4	5	1	2	3	4	5	2	5
Action	sd	sd	sd	sd	sd	su	su	su	su	su	hw	hw

Table 5. Tasks in presence of an object.

Task	**1**	**2**	**3**	**4**	**5**	**6 ***
Object location	Figure 4a	Figure 4b	Figure 4c	Figure 4d	Figure 4e	Figure 4f
Walk Forward	A to a	C to c	A to c	C to a	B to b	hw
Repetitions	12	12	12	12	12	12
Walk Reverse	a to A	c to C	a to C	c to A	b to B	su/sd
Repetitions	12	12	12	12	12	12/12

* In Task 6, the robot did not go forward or reverse, but performed hand-wave, stand-up and sit down action, where each action was repeated 12 times.

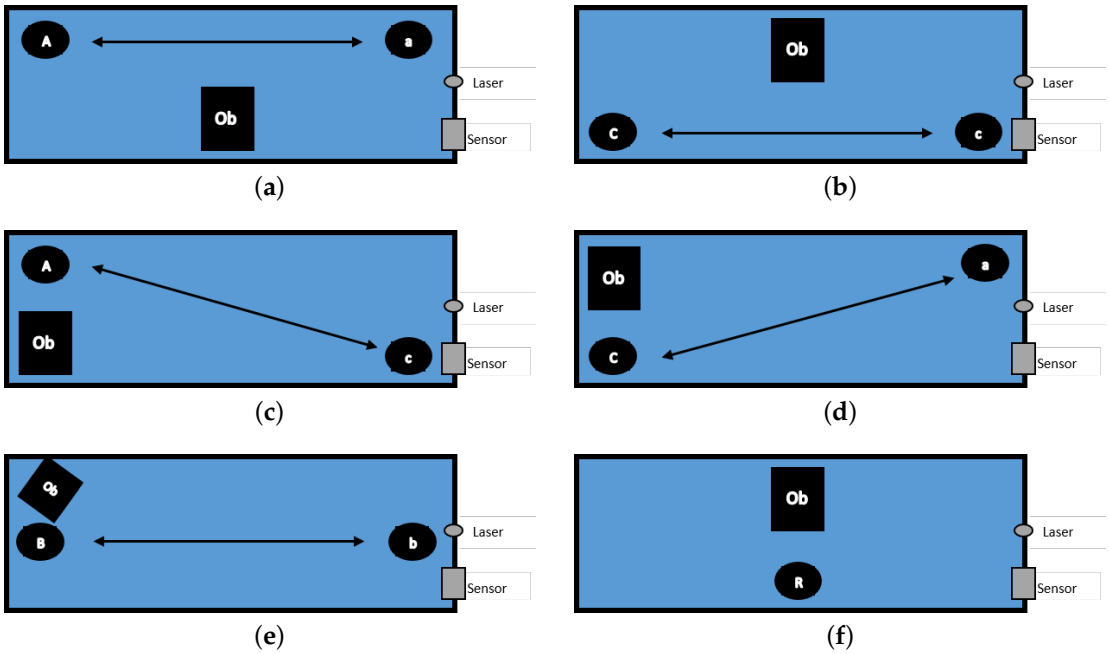


Figure 4. Positions of object during various robot actions.

3.2. Two-Robot Setup

We also devised new setup with two NAO V4 humanoid robots. Firstly, one robot was standing still at Position 1 and the other robot at Position 2 walks forward and reverse to Positions 3 and 4, as shown in Figure 5. This action was repeated 10 times. In the next experiment, both robots performed actions simultaneously. Performed actions are listed in Table 6).

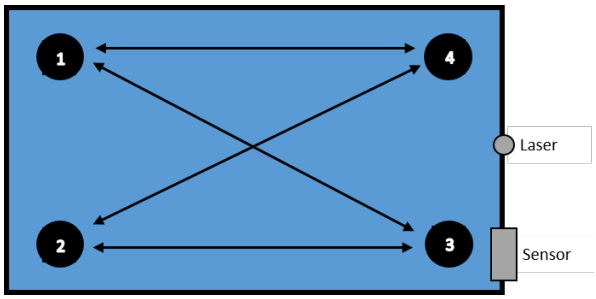


Figure 5. Position of two robots during actions.

Table 6. Actions performed by two robots.

Repetition	Robot 1 Action	Position	Robot 2 Action	Position
10	Forward	1 to 4	Forward	2 to 3
10	Sit Down	1	Sit Down	2
10	Stand Up	1	Stand Up	2
10	Sit Down	3	Sit Down	4
10	Stand Up	3	Stand Up	4
10	Hand-Wave	1	Hand-Wave	2
10	Hand-Wave	3	Hand-Wave	4
10	Stand	1	Forward	2 to 3
10	Stand	1	Forward	2 to 4

In Figure 6, we illustrate a few examples of preprocessed data, which was used in training. Columns correspond to different actions and the rows are different examples. That is, the sequences consisted of a time series of 500-dimensional vectors.

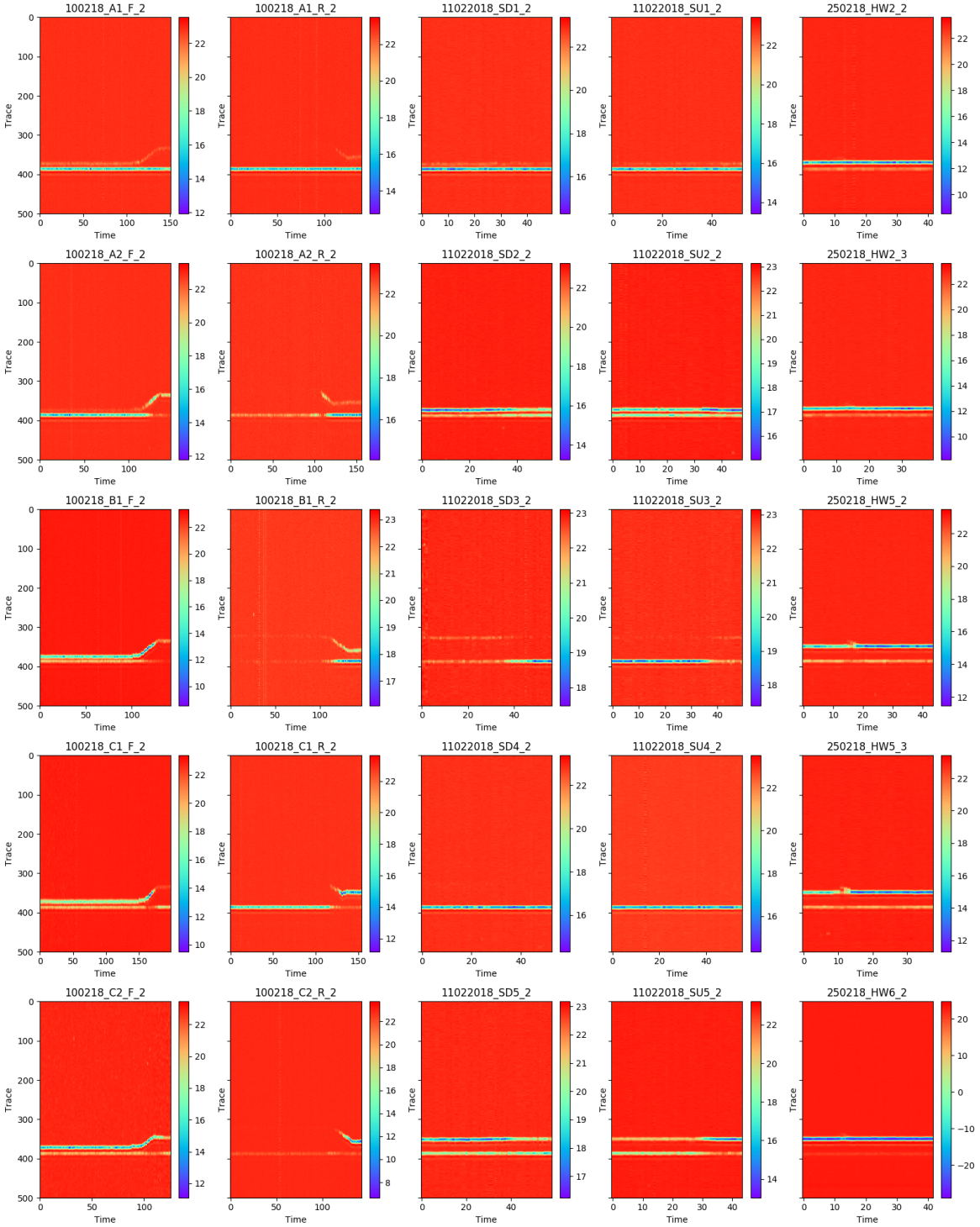


Figure 6. Visualisation of the traces throughout time (x-axis). Columns correspond to different actions (respectively, forward walking, reverse walking, sitting down, standing up, and waving), whereas rows correspond to different examples. The titles on the subplots correspond to the sequence files in the dataset.

4. Method

We chose a recurrent neural network (RNN) as our baseline. Recurrent nets are able to model multivariate time-series—in our case, time-of-flight measurements—and output a class prediction by considering the whole temporal sequence. In particular, our choice was a RNN with Gated-Recurrent Unit (GRU) cells. These cells can retain long-temporal information using internal gates and a set of optimisable parameters.

4.1. Gated-Recurrent Unit

We briefly introduce GRUs following the notation from [53]. Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$, $\mathbf{x}_t \in \mathbb{R}^n$ be a sequence of T observations and $y \in C$ its ground truth class label. At each time step t, a GRU cell receives $\mathbf{x}_t$ and outputs an activation $h_t \in \mathbb{R}^m$ response

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j \tilde{h}_t^j \tag{1}$$

by combining activation at previous time step h_{t-1}^j and a candidate activation from the current time step $\tilde{h}_t^j$.

The trade-off factor z_t^j, namely *update gate*, is calculated as

$$z_t^j = \sigma(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1})^j, \tag{2}$$

where $W_z \in \mathbb{R}^{m \times n}$ and $U_z \in \mathbb{R}^{m \times m}$ are optimisable parameters shared across all t and σ a sigmoid function that outputs values in the interval (0,1).

In its turn, the *candidate activation* is calculated

$$\tilde{h}_t^j = \tanh(W \mathbf{x}_t + U(\mathbf{r}_t \odot \mathbf{h}_{t-1}))^j, \tag{3}$$

where $\odot$ is the element-wise product of two vectors and $\mathbf{r}_t$ also known as *reset gate*. Note that $W \in \mathbb{R}^{m \times n}$ and $U \in \mathbb{R}^{m \times m}$ are different sets of parameters from W_z and U_z.

Similar to the update gate $\mathbf{z}_t$, the *reset gate* is

$$r_t^j = \sigma(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1})^j. \tag{4}$$

Finally, the last GRU activation at time T is input to a dense layer with softmax activation function. From the dense layer, the logit value z^i is computed by

$$z^i = \sum_j w_s^{ij} h_T^j, \tag{5}$$

where $W_s = (w_s^{ij})$ are the softmax layer weights. Then, the softmax activation function can be applied to output the sequence classification label

$$\hat{y}^i = \frac{e^{z^i}}{\sum_i e^{z_i}}. \tag{6}$$

4.2. Bidirectional GRU and Stacked Layers

Bidirectional recurrent networks consist of two independent networks processing the temporal information in the two temporal dimensions, forward and reverse, so their activation outputs are concatenated. The input of the reverse recurrent network is simply the reversed input sequence. The logit value computation becomes

$$z^i = \sum_j w_s^{ij} [h^j_{\text{fw},T}, h^j_{\text{rv},0}], \tag{7}$$

where $[\cdot, \cdot]$ is the concatenation of forward and reverse GRUs activations.

In addition, GRU layers can be stacked to form a deeper GRU architecture. The first GRU layer receives as input the sequence of observations $\mathbf{x}$, whereas each subsequent layers are fed with activation outputs from the previous layer. We finally apply the softmax dense layer to the activations of the deepest stacked layer.

4.3. Baseline

Our architecture is a two-layer bidirectional GRU, each GRU with 512 neurons (experimentally chosen). The size of the softmax dense layer is the number of classes $|C|$. Figure 7 illustrates the architecture.

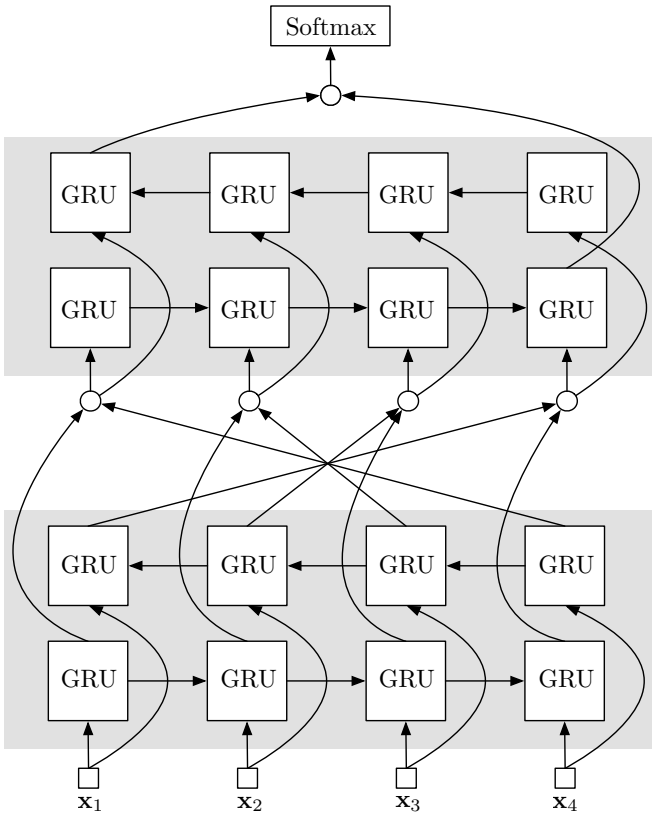


Figure 7. The two-layer bidirectional GRU baseline architecture. Arrays represent information flow, grey rectangles are bidirectional GRU layers, and circles represent the concatenation operation.

5. Experimental Results and Discussion

5.1. Learning Model Details and Code Implementation

Among different RNN cells, we chose Gated-Recurrent Units (GRU) for our baseline architecture. Compared to other recurrent cells, such as Long-Term Short Memory (LSTM) cells, these require a reduced number of parameters while still retaining long-term temporal information and providing highly competitive performance [54]. GRU is also often chosen over LSTM because hidden states are fully exposed and hence easier to interpret.

For the model computations, we entirely relied on GPU programming. In particular, our implementation is based on Keras [55], a GPU-capable deep-learning library written in Python. As for the GPU device itself, we utilised an NVIDIA Titan Xp with 12 GB of GDDR5X memory.

5.2. Ablation Experiments on GRU Architectures

To determine the best GRU architecture, we first performed a set of binary classification experiments on the following actions: forward (walking), reverse (walking), sit-down, standing up, and handwaving. We report the performance in terms of accuracy (averaging accuracies over a 10-fold cross validation). In Table 7, we illustrate the ablation experiments on different multi-layer and bidirectional GRU architectures with fixed hidden layer size to 64 neurons. For each architecture and target action, we trained a different GRU model for 25 epochs, which was enough to avoid under-fitting in the most complex model (two-layer biGRU).

In particular, the most complex model, two-layer biGRU, was the one that provided the best result. This showed how both multiple and bidirectional layers can help to model single-pixel time-of-light data sequences. In particular, adding a second stacked layer provided a +5.09% improvement over one single layer, whereas the bidirectionality increased accuracy by 4.4%. The +8.37% gain from using both showed how those two architecture variations are highly complementary when dealing with our data.

Table 7. Comparison on GRU models with multiple layers and/or bidirectionality. In this ablation, we defined a set of five binary problems: forward, reverse, sit-down, stand-up, and hand-wave actions. The results reported are class-weighted accuracies averaged over a 10-fold cross validation. The "Average" column is the average of performances on binary problems.

	Forward	**Reverse**	**Sit-Down**	**Stand-Up**	**Handwave**	**Average**
GRU (1-layer, 64-hidden)	87.28	83.85	76.48	78.15	0.945	84.05
GRU (two-layer, 64-hidden)	88.48	90.06	85.75	86.41	94.99	89.14
biGRU (1-layer, 64-hidden)	89.28	87.62	82.49	86.85	96.02	88.45
biGRU (two-layer, 64-hidden)	91.42	91.08	90.07	92.51	97.01	92.42

Next, using a two-layer biGRU, we performed another set of ablation experiments on hidden layer sizes: $\{32, 64, 128, 256, 512\}$. Since the hidden layer size drastically affects the number of parameters to optimise during the training stage, each model was trained during a different number of epochs: $\{10, 25, 50, 100, 200, 400\}$, respectively. Results are shown in Table 8.

The largest model, i.e., 512 hidden layer neurons, performed the best. Its +5.62% gain with respect to the smallest two-layer biGRU model with 32 neurons demonstrated room for improvement from using more complex models despite the presumed simplicity of single-pixel time-of-flight time-series. However, we discarded further increasing the hidden size because of computational constraints: enlarging the hidden layer causes an exponential grow of the number of parameters to train. In particular, a model with 32 hidden neurons consisted of 121 K parameters, whereas 512 hidden neurons increased the size up to 7.8 M (and 37.7 M in the case of 1024 neurons); this and the saturation of accuracy discouraged us to keep enlarging the hidden layer size.

Before further experimentation with GRU recurrent nets, we compared the best performing model to its analogous LSTM variant (two-layer biLSTM with 512 hidden layer neurons). In Table 9, we show how GRU could obtain competitive performance with LSTM. The marginal improvement of 0.56% obtained by LSTM requires a substantial increment of the number of parameters, especially when considering larger models. In the case of 512 hidden size, LSTM has 2.6M additional parameters to optimise when compared to the GRU version. For further experiments, we stuck to the biGRU (two-layer, 512 hidden neurons) architecture.

Table 8. Hidden layer size experiments on five binary problems (see Columns 2–6). The results reported are class-weighted accuracies averaged over a 10-fold cross validation. The "Average" column is the average of performances on binary problems.

	Forward	Reverse	Sit-Down	Stand-Up	Handwave	Average
biGRU (two-layer, 32-hidden)	89.97	89.78	87.89	89.47	95.92	90.61
biGRU (two-layer, 64-hidden)	91.42	91.08	90.07	92.51	97.01	92.42
biGRU (two-layer, 128-hidden)	93.10	93.63	92.98	95.32	97.70	94.55
biGRU (two-layer, 256-hidden)	93.76	94.17	94.34	95.52	98.89	95.34
biGRU (two-layer, 512-hidden)	94.94	95.20	95.02	96.70	99.29	96.23

Table 9. GRU versus LSTM on 5 binary problems (see Columns 2–6). The results reported are class-weighted accuracies averaged over a 10-fold cross validation. The "Average" column is the average of performances on binary problems.

	Forward	Reverse	Sit-Down	Stand-Up	Handwave	Average
biGRU (two-layer, 64-hidden)	91.42	91.08	90.07	92.51	97.01	92.42
biLSTM (two-layer, 64-hidden)	91.56	96.91	92.02	89.84	94.58	92.98

5.3. Final Experiments

After having fixed the final GRU model architecture to tow bidirectional stacked layers with 512 hidden neurons, we performed evaluated its performance in multiclass classification and also other experiments to ensure the generalisation capabilities of our approach.

5.3.1. Multiclass Classification

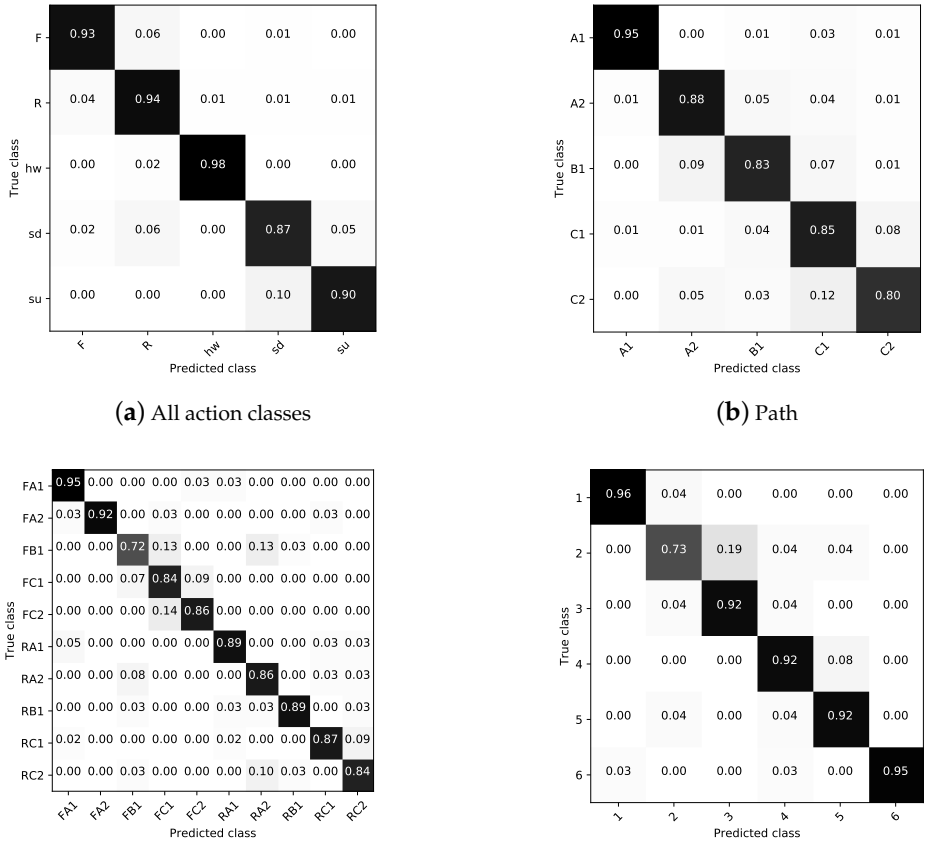
To evaluate the missclassifications and potential confusion among classes from our previous binary problems, we first defined a multiclass problem with labels those same labels: {F, R, sd, su, hw}, where F is forward, R is reverse, sd is sit down, su is stand up, and hw is hand-wave. In this five-class problem, the model was able to correctly predict 92.67% of actions (see column "Actions" in Table 10). As shown in Figure 8a, the confusion is introduced by the semantically similar classes, either forward and reverse or sit-down and stand-up. The hand-wave classification was almost perfect, only confused once as a reverse instance in 50 hw examples.

The second and third experiments were intended to classify the walking path. The former was not distinguishing walking direction. We hence defined two separate sets of labels, {A1, A2, B1, C1, C2} and {FA1, FA2, FB1, FC1, FC2, RA1, RA2, RB1, RC1, RC2}, respectively, where letters (F) and (R) before action label are used to distinguish between action in forward or reverse direction. As shown in Table 10, the model performed similarly in the two cases, with slightly worse performance not considering the walking direction (86.23%) than when doing so (86.65%). Figure 8 shows the confusion matrices for these two experiments.

Finally, in the fourth and last experiment, we labelled the setup in which the action was occurring with labels $\{1, \ldots, 6\}$, which correspond to tasks listed in Table 5. In this experiment, the robot had to perform various tasks, in addition to performing an action an object is also present in the same environment. Its location and performed actions can be seen in Figure 4. The accuracies obtained from those are summarised in Table 10, while Figure 8 illustrates class confusions.

Table 10. Classification on four multiclass problems obtained by biGRU (two-layer, 512-hidden) baseline. The results reported are class-weighted accuracies averaged over a 10-fold cross validation.

Actions	Path	Directed-Path	Setup
92.67	86.23	86.65	90.00



(**a**) All action classes

(**b**) Path

(**c**) Directed path

(**d**) Setup

Figure 8. Confusion matrices (row-wise normalised) from multiclass classification experiments from Table 10.

5.3.2. Model Generalisation on Actions and Two Robots

In this section, we evaluate the generalisation capabilities of the models when learning from single-pixel time-of-flight patterns.

Each action was captured a certain amount of repetitions. During this repetition, the path in walking actions (forward and reverse) or initial position (sit-down, stand-up, and hand-wave) were varied. In this experiment, we wanted to take this into account and try to learn by excluding from training the all the repetitions of one action to assert the model is not overfitting due to repetitions being very similar patterns. For that, we changed our validation procedure to leave-one-rep set-out, i.e., we predicted a repetition set all at once in the test set, and did not use repetitions from the same respect during training. Results are presented in Table 11. If we compare to those to results from the same model, i.e., biGRU (two-layer, 512-hidden), in Table 10, we can observe there was no drop in accuracy, but a slight improvement—probably due to both the generalisation capabilities and the fact that we could use more data to train across folds.

Table 11. Leave-one-rep set-out cross-validation (LOROCV) experiment using biGRU (two-layer, 512-hidden). These are the same as those from last row in Table 8, but using LOROCV instead of 10-fold CV.

	F	**R**	**sd**	**su**	**hw**	**Average**
biGRU (two-layer, 512-hidden, 10fCV)	94.94	95.20	95.02	96.70	99.29	96.23
biGRU (two-layer, 512-hidden, LOROCV)	96.77	95.14	93.36	97.11	100.0	96.47

All sequences were with just one robot performing actions. A separate set of sequences was used to test action classification when two robots were present, as shown in Table 12. These sequences were only used in the test phase (only one-robot sequences were used for training). In particular, we analysed three different scenarios: (1) one robot acted, while the other one stood still; (2) the two robots performed the same action; and (3) each robot performed a different action.

From results in Scenario (1), we observed the standing-up robot did not interfere in the other action category prediction. In fact, the model failed to predict stand-up action since the other actions presented a more dominant motion pattern that interfere in the stand-up pattern learned from one-robot actions.

Table 12. Two-robot experiments in three different scenarios: one robot standing up while other performing a particular action, the two robots performing the same action, and the two performing each a different action. Each scenario is a separate test set with a different number of examples. In brackets, the number of positive examples for each class in each scenario. Since positive/negative classes are, we report class-weighted accuracies (%).

	#{Examples}	F	R	sd	su	hw
One robot standing up and sitting down	100	80.00 (50)	88.00 (50)	95.00 (0)	15.00 (100)	100.00 (0)
Same two actions	70	25.00 (10)	72.86 (10)	75.00 (20)	54.00 (20)	50.00 (20)
Two different actions	20	50.00 (10)	100.00 (0)	95.00 (10)	55.00 (10)	55.00 (10)

5.4. Discussion

In this paper, we propose a concept for detection actions while preserving the test subjects (NAO V4 robot) privacy. Our concept relies on recording only the temporal evolution of light pulses scattered back from the scene. Such data trace to record one action contains sequence of one-dimensional arrays of voltage values acquired by the single-pixel detector after amplifying and detection by the data acquisition system at 6 GHz repetition rate. The data trace is very compact and easy to process, compared to videos, containing sequences of 2D images.

The data volume reduction is achieved by controlled illumination and single pixel detector without any spatial resolution. The scene was illuminated with a diverging, speckled light pulse of 30 picosecond (30×10^{-12} ps) duration. The method would also work in different scenes, where most of the objects are static.

Compared to 2D images, hardly any information about the colours, object, their shapes and positions could be retrieved from the data traces by classical method. Although quite similar to the neural networks, a human can distinguish the actions and perhaps also clearly differentiate moving directions from the data traces.

The research in hand clearly articulates the core properties of movement—it imprints a temporal evolution to even most simple data trace. Owing to the interdisciplinary approach through combining the tools of photonics (modern, application oriented optics and light detection) and computer science, one is capable of reducing the data rate. The result has high potential to provide cost effective surveillance systems to aid societies to look after of public order, and take care of young, elderly and injured members.

The photonics and data acquisition schemes used in this experiment are unlikely to become widespread owing to their high cost and other features. However, detectors and laser systems capable of providing suitable illumination and detection properties in affordable price range are being developed and will enter the market in near future.

6. Conclusions

This research work proposed a new methodology for action recognition while preserving the test subjects privacy. The proposed method uses only the temporal evolution of light pulses scattered back from the scene. Advanced machine learning algorithms, namely RNN and LSTM, were adopted for data analysis and demonstrated successful action recognition. The experimental results show that our proposed method could achieve high recognition rate for five actions, namely walking forward, walking reverse, sitting down, standing up, and waving hand, with an average recognition rate of 96.47%. In this work, we additionally studied action recognition when multiple concurrent actors are present in the scene.

In future work, we will conduct further experiments, including more complex actions, such as running, jumping, and head movements. We are planning to record higher number of samples to conduct a better generalisation capabilities of our proposed approach.

Author Contributions: Conceptualization, H.V.-L., S.E., C.O. and G.A.; Data curation, I.O., A.H., K.M.P., S.-M.V. and S.O.; Funding acquisition, G.A.; Investigation, H.V.-L., S.O., S.E. and G.A.; Methodology, A.C., E.A., A.V., H.V.-L., S.E. and G.A.; Software, A.C.; Writing—original draft, A.H. and G.A.; Writing—review & editing, I.O., E.A., S.-M.V., H.V.-L., S.O., S.E., C.O. and G.A.

Funding: This work was partially supported by Estonian Research Council Grants (PUT638, PUT1075, PUT1081), The Scientific and Technological Research Council of Turkey (TÜBITAK) (Project 1001-116E097), the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund, the Spanish Project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya. This project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 665919. This work was partially supported by ICREA under the ICREA Academia programme.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp and V GPUs used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fernando, B.; Gavves, E.; Oramas, J.M.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.
2. Nasrollahi, K.; Escalera, S.; Rasti, P.; Anbarjafari, G.; Baro, X.; Escalante, H.J.; Moeslund, T.B. Deep learning based super-resolution for improved action recognition. In Proceedings of the IEEE 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), Orleans, France, 10–13 November 2015; pp. 67–72.
3. Haque, M.A.; Bautista, R.B.; Noroozi, F.; Kulkarni, K.; Laursen, C.B.; Irani, R.; Bellantonio, M.; Escalera, S.; Anbarjafari, G.; Nasrollahi, K.; et al. Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 250–257.
4. Ponce-López, V.; Escalante, H.J.; Escalera, S.; Baró, X. Gesture and Action Recognition by Evolved Dynamic Subgestures. In Proceedings of the BMVC, Swansea, UK, 7–10 September 2015; pp. 129.1–129.13.
5. Wan, J.; Escalera, S.; Anbarjafari, G.; Escalante, H.J.; Baró, X.; Guyon, I.; Madadi, M.; Allik, J.; Gorbova, J.; Lin, C.; et al. Results and Analysis of ChaLearn LAP Multi-modal Isolated and Continuous Gesture Recognition, and Real Versus Fake Expressed Emotions Challenges. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3189–3197.
6. Corneanu, C.; Noroozi, F.; Kaminska, D.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on Emotional Body Gesture Recognition. *IEEE Trans. Affect. Comput.* **2018**. [CrossRef]
7. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1473. [CrossRef]

8. Jahromi, M.N.; Bonderup, M.B.; Asadi-Aghbolaghi, M.; Avots, E.; Nasrollahi, K.; Escalera, S.; Kasaei, S.; Moeslund, T.B.; Anbarjafari, G. Automatic Access Control Based on Face and Hand Biometrics in a Non-Cooperative Context. In Proceedings of the 2018 IEEE Winter Applications of Computer Vision Workshops (WACVW), Lake Tahoe, NV, USA, 15 March 2018; pp. 28–36.
9. Sapiński, T.; Kamińska, D.; Pelikant, A.; Ozcinar, C.; Avots, E.; Anbarjafari, G. Multimodal Database of Emotional Speech, Video and Gestures. In Proceedings of the International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 153–163.
10. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
11. Lusi, I.; Junior, J.C.J.; Gorbova, J.; Baró, X.; Escalera, S.; Demirel, H.; Allik, J.; Ozcinar, C.; Anbarjafari, G. Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 809–813.
12. Avots, E.; Sapiński, T.; Bachmann, M.; Kamińska, D. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **2018**, 1–11. [CrossRef]
13. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Fusion of classifier predictions for audio-visual emotion recognition. In Proceedings of the IEEE 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 61–66.
14. Guo, J.; Lei, Z.; Wan, J.; Avots, E.; Hajarolasvadi, N.; Knyazev, B.; Kuharenko, A.; Junior, J.C.S.J.; Baró, X.; Demirel, H.; et al. Dominant and Complementary Emotion Recognition From Still Images of Faces. *IEEE Access* **2018**, *6*, 26391–26403. [CrossRef]
15. Grobova, J.; Colovic, M.; Marjanovic, M.; Njegus, A.; Demire, H.; Anbarjafari, G. Automatic hidden sadness detection using micro-expressions. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 828–832.
16. Kulkarni, K.; Corneanu, C.; Ofodile, I.; Escalera, S.; Baró, X.; Hyniewska, S.; Allik, J.; Anbarjafari, G. Automatic recognition of facial displays of unfelt emotions. *IEEE Trans. Affect. Comput.* **2018**. [CrossRef]
17. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the BMVC, Swansea, UK, 7–10 September 2015; Volume 1, p. 6.
18. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
19. Haamer, R.E.; Kulkarni, K.; Imanpour, N.; Haque, M.A.; Avots, E.; Breisch, M.; Nasrollahi, K.; Escalera, S.; Ozcinar, C.; Baro, X.; et al. Changes in facial expression as biometric: A database and benchmarks of identification. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 621–628.
20. Tertychnyi, P.; Ozcinar, C.; Anbarjafari, G. Low-quality fingerprint classification using deep neural network. *IET Biom.* **2018**, *7*, 550–556. [CrossRef]
21. Zhang, C.L.; Zhang, H.; Wei, X.S.; Wu, J. Deep bimodal regression for apparent personality analysis. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 August 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 311–324.
22. Gorbova, J.; Avots, E.; Lüsi, I.; Fishel, M.; Escalera, S.; Anbarjafari, G. Integrating Vision and Language for First-Impression Personality Analysis. *IEEE MultiMedia* **2018**, *25*, 24–33. [CrossRef]
23. Yang, J.; Nguyen, M.N.; San, P.P.; Li, X.; Krishnaswamy, S. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; Volume 15, pp. 3995–4001.
24. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 1894–1903.
25. Ordóñez, F.J.; Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [CrossRef]

26. Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y.; Wang, Y. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors* **2017**, *17*, 818. [CrossRef]
27. Kirmani, A.; Hutchison, T.; Davis, J.; Raskar, R. Looking around the corner using transient imaging. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 159–166.
28. Velten, A.; Willwacher, T.; Gupta, O.; Veeraraghavan, A.; Bawendi, M.G.; Raskar, R. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nat. Commun.* **2012**, *3*, 745. [CrossRef]
29. Buttafava, M.; Zeman, J.; Tosi, A.; Eliceiri, K.; Velten, A. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Opt. Express* **2015**, *23*, 20997–21011. [CrossRef]
30. Besl, P.J. Active optical range imaging sensors. In *Advances in Machine Vision*; Springer: Berlin/Heidelberg, Germany, 1989; pp. 1–63.
31. Antipa, N.; Kuo, G.; Heckel, R.; Mildenhall, B.; Bostan, E.; Ng, R.; Waller, L. DiffuserCam: Lensless single-exposure 3D imaging. *Optica* **2018**, *5*, 1–9. [CrossRef]
32. Gatti, A.; Brambilla, E.; Bache, M.; Lugiato, L.A. Ghost imaging with thermal light: Comparing entanglement and classicalcorrelation. *Phys. Rev. Lett.* **2004**, *93*, 093602. [CrossRef]
33. Shapiro, J.H. Computational ghost imaging. *Phys. Rev.* **2008**, *78*, 061802. [CrossRef]
34. Sun, M.J.; Edgar, M.P.; Gibson, G.M.; Sun, B.; Radwell, N.; Lamb, R.; Padgett, M.J. Single-pixel three-dimensional imaging with time-based depth resolution. *Nat. Commun.* **2016**, *7*, 12010. [CrossRef]
35. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3d points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
36. Caramazza, P.; Boccolini, A.; Buschek, D.; Hullin, M.; Higham, C.; Henderson, R.; Murray-Smith, R.; Faccio, D. Neural network identification of people hidden from view with a single-pixel, single-photon detector. *arXiv* **2017**, arXiv:1709.07244.
37. Sanchez-Riera, J.; Čech, J.; Horaud, R. Action recognition robust to background clutter by using stereo vision. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 332–341.
38. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
39. Papadopoulos, G.T.; Axenopoulos, A.; Daras, P. Real-time skeleton-tracking-based human action recognition using kinect data. In Proceedings of the International Conference on Multimedia Modeling, Dublin, Ireland, 6–10 January 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 473–483.
40. Fofi, D.; Sliwa, T.; Voisin, Y. A comparative survey on invisible structured light. In *Machine Vision Applications in Industrial Inspection XII*; International Society for Optics and Photonics: San Diego, CA, USA, 2004; Volume 5303, pp. 90–99.
41. Smisek, J.; Jancosek, M.; Pajdla, T. 3D with Kinect. In *Consumer Depth Cameras for Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 3–25.
42. Faccio, D.; Velten, A. A trillion frames per second: The techniques and applications of light-in-flight photography. *Rep. Prog. Phys.* **2018**, *81*, 105901. [CrossRef] [PubMed]
43. Pandharkar, R.; Velten, A.; Bardagjy, A.; Lawson, E.; Bawendi, M.; Raskar, R. Estimating motion and size of moving non-line-of-sight objects in cluttered environments. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 265–272.
44. Heide, F.; Hullin, M.B.; Gregson, J.; Heidrich, W. Low-budget transient imaging using photonic mixer devices. *ACM Trans. Graph. (ToG)* **2013**, *32*, 45. [CrossRef]
45. Gariepy, G.; Tonolini, F.; Henderson, R.; Leach, J.; Faccio, D. Detection and tracking of moving objects hidden from view. *Nat. Photonics* **2016**, *10*, 23–26. [CrossRef]
46. Warburton, R.E.; Chan, S.; Gariepy, G.; Altmann, Y.; McLaughlin, S.; Leach, J.; Faccio, D. Real-Time Tracking of Hidden Objects with Single-Pixel Detectors. In *Imaging Systems and Applications*; Optical Society of America: San Diego, CA, USA, 2016; p. IT4E–2.
47. Chan, S.; Warburton, R.E.; Gariepy, G.; Leach, J.; Faccio, D. Non-line-of-sight tracking of people at long range. *Opt. Express* **2017**, *25*, 10109–10117. [CrossRef]
48. Jia, L.; Radke, R.J. Using time-of-flight measurements for privacy-preserving tracking in a smart room. *IEEE Trans. Ind. Inform.* **2014**, *10*, 689–696. [CrossRef]

49. Tao, S.; Kudo, M.; Nonaka, H. Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network. *Sensors* **2012**, *12*, 16920–16936. [CrossRef] [PubMed]
50. Kawashima, T.; Kawanishi, Y.; Ide, I.; Murase, H.; Deguchi, D.; Aizawa, T.; Kawade, M. Action recognition from extremely low-resolution thermal image sequence. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
51. Dai, J.; Saghafi, B.; Wu, J.; Konrad, J.; Ishwar, P. Towards privacy-preserving recognition of human activities. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4238–4242.
52. Xu, M.; Sharghi, A.; Chen, X.; Crandall, D.J. Fully-Coupled Two-Stream Spatiotemporal Networks for Extremely Low Resolution Action Recognition. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1607–1615.
53. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
54. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
55. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 4 February 2019).